CALIBRATION OF TRIP DISTRIBUTION BY GENERALISED LINEAR MODELS

John Shrewsbury

A thesis submitted in partial fulfilment of the requirements for the Degree of

Doctor of Philosophy in Transportation Engineering

Department of Civil and Natural Resources Engineering

University of Canterbury

2012

Acknowledgements

This research was funded by the New Zealand Transport Agency (NZTA), and is also published on its website at www.nzta.govt.nz/resources/research/reports/473/index.html as an NZTA research report. The research is independent, and should not be regarded as the opinion, responsibility or policy of the NZTA or indeed any agency of the NZ Government. I am grateful to the NZTA for their financial support and temporal tolerance.

I would also like to express my gratitude to the many other people and organisations who have helped and encouraged this research, notably Professor Ben Heydecker of University College London who gave generously of his time, mathematical insight, and enthusiasm for probing the boundaries of knowledge while I was in the UK. In New Zealand, my academic supervisor Professor Alan Nicholson and many other members of the University of Canterbury supported my studies, and I drew a great depth of experience in both countries from NZTA's peer reviewers, Tony Brennand and Professor Howard Kirby. Greater Wellington regional council provided the Wellington Transport Strategy Model with many supporting datasets, and Citilabs supplied the MVESTM software with source code and its Cube Voyager modelling environment. The late Professor John Nelder of Imperial College and his colleagues at VSN, who implement Genstat, gave many insights into the statistical scope and computational complexities of GLMs and HGLMs. One of the most telling of these was John Nelder's comment on an arcane point, 'Yes, I wondered that too'.

Abstract

Generalised linear models (GLMs) provide a flexible and sound basis for calibrating gravity models for trip distribution, for a wide range of deterrence functions (from steps to splines), with K factors and geographic segmentation. The Tanner function fitted Wellington Transport Strategy Model data as well as more complex functions and was insensitive to the formulation of intrazonal and external costs. Weighting from variable expansion factors and interpretation of the deviance under sparsity are addressed.

An observed trip matrix is disaggregated and fitted at the household, person and trip levels with consistent results. Hierarchical GLMs (HGLMs) are formulated to fit mixed logit models, but were unable to reproduce the coefficients of simple nested logit models.

Geospatial analysis by HGLM showed no evidence of spatial error patterns, either as random K factors or as correlations between them. Equivalence with hierarchical mode choice, duality with trip distribution, regularisation, lorelograms, and the modifiable areal unit problem are considered.

Trip distribution is calibrated from aggregate data by the MVESTM matrix estimation package, incorporating period and direction factors in the intercepts. Counts across four screenlines showed a significance similar to a thousand-household travel survey. Calibration was possible only in conjuction with trip end data. Criteria for validation against screenline counts were met, but only if allowance was made for error in the trip end data.

General terminology

The terms trip distribution, destination choice and gravity model are used interchangeably. A simple form is:

$$t_{ij} = P_i A_j p_i a_j \exp(-\lambda C_{ij})$$

where t_{ij} are the trips for the movement from production zone i to attraction zone j

P_i and A_j are the production and attraction trip-end totals

p_i and a_j are the production and attraction balancing factors

λ is the cost coefficient for Exponential deterrence [or γ for a Power function, $f(C) = C^{-\gamma}$]

C_{ij} is the cost of travel from zone i to zone j

[k is the Index for cost bands in empirical trip distribution or screenlines in matrix estimation]

The usual effect of cost is to deter travel. An explicit negative sign is included in the cost term of deterrence functions

$$f(C) = \exp(-\lambda C) \text{ or } C^{-\gamma}$$

so that coefficients of cost λ or its logarithm γ are positive in this usual case, and larger coefficients imply a greater influence of cost. Costs are in generalised minutes.

Sectors are groupings of zones – a part of the study area.
Segments are groupings of production-attraction (PA) or origin-destination (OD) movements– a part of the matrix. A segment may be the intersection of a production sector with an attraction sector.

K factors are constants in the deterrence function that vary between segments.
L factors are coefficients of cost in the deterrence function that vary between segments.

Productions (P) and **attractions** (A) are the home and non-home end of the trip, relating travel to land use
Origin (O) and **destination** (D) are the start and end of the trip.

Empty zones have no observed trip ends, forming complete rows or columns of zeros in an observed trip matrix. They do not contribute any information about the distribution of trips.
Zero cells are matrix cells whose movements have been observed, but for which the observation is zero. This is useful information, that the volume of the movement is probably small.
Null cells are matrix cells whose movements cannot be observed. They contain no information.

Flat matrices have cells simply proportional to their trip ends, without any effect of cost. They are the null model for trip distribution.

Aggregate and **disaggregate** are relative to model (WTSM) zoning.

Logarithms are natural, to the base e=2.718…, and significances are at the 5% level unless stated otherwise.

The principal dataset was a fully-observed 24-hour weekday internal matrix of home-based work (**HBW**) person-trips by car from the 2001 household interview survey (**HIS**) for the Wellington Transport Stratetegy Model (**WTSM**). External trip were observed by roadside interview survey (**RSI**).

These terms and conventions have been followed as far as practicable in this thesis. In referring to other works, and particularly when summarising them or abstracting key features in brief, the original terminology is retained for ease of reference to the original work. There is an introduction to transport modelling in section 1.2, to trip distribution in 1.3, and to generalised linear models in 3.2. Appendix J provides a glossary.

Contents

List of figures

List of tables

1 Introduction

This thesis describes the application of the statistical methods of generalised linear models to the calibration of trip distribution in the context of a major working transport model. While the potential of this method of calibration has been recognised academically for some time, there has not been sufficient computing power to apply it to practical models until recently. There is thus little experience of its use in practice. In the meantime, the statistical methods have been extended, in particular to hierarchical GLMs (HGLMs).

1.1 Outline

1.1.1 Trip distribution

The evaluation of any substantial scheme in a complex transport network requires a detailed knowledge of the demand for travel. In particular, the origins and destinations of trips need to be known to calculate re-routing. This information is hard to collect thoroughly: screenline surveys miss trips that do not cross them, and household or workplace surveys are too expensive to give a good sample of all trips. However, both can be used to calibrate models that link origins to destinations. These trip distribution models have the advantage that they can predict travel patterns for different transport systems in the future, as well as land-use changes which can sometimes be addressed by simpler factoring.

1.1.2 Calibration

The calibration of such models is a complex process. Although analytical methods to find the best fitting parameters have been available for some time, they are not implemented in all modelling packages, and trial-and-error methods are still employed. Even the best-fitting models can leave much to be desired and ad hoc adjustments, by 'K' factors, are often introduced.

1.1.3 Statistical approach with generalised linear models

The analytical calibration process can be treated as an advanced form of regression called a generalised linear model (GLM). This is based on likelihood maximisation, and provides statistical measures of the fit of models and the significance and accuracy of their components. The basic methodology is well established and mature, having been developed over a period of 30 years. There have been several more recent advances, in particular mixed modelling which allows different sources of error to be distinguished within the model.

1.1.4 Wellington travel data and model structure

The Wellington Transport Strategy Model (WTSM) was substantially rebuilt around a major household travel survey concurrent with the 2001 Census. Greater Wellington Regional Council (GW) kindly made the model and its data available for the research. The survey recorded the travel patterns of 2538 households, and had already been prepared for model calibration. The WTSM also provided the sound framework of a fully developed model, while presenting the practicalities of a working model to be addressed in the research. The objective of the study was not to re-evaluate the WTSM, but to take it as a single full-scale example on which the alternative approaches to calibration offered by GLMs could be developed and assessed. Commuting (home-based work (HBW)) trips by car were taken as the prime example for study.

1.1.5 Support for other major transport models

The study was funded by the New Zealand Transport Agency (NZTA), and was intended to support the redevelopment of the Auckland and Christchurch transport models around the 2006 Census. A proposal was made to apply the methodology as a parallel stream within the development of these models, but was rejected. When it became apparent that the methodology was not being adopted for the development of these models, the NZTA agreed that the remainder of the study should explore the wider potential of the methodology, rather than package it for practical application.

1.1.6 Land-use formulation and disaggregation

A land-use model was commissioned in parallel with the Auckland transport model. Two of the three final tenders for the Auckland models proposed to predict the distribution of commuting from the land-use model rather than from the transport model. Although these proposals were not accepted, the WTSM household interview survey was re-formulated to a land-use viewpoint of housing and employment, rather than a transport viewpoint of trip productions and attractions. This led to a major change in the scale of weighting, and to disaggregation from zones to households, persons and trips.

1.1.7 Distribution and mode choice – mixed logit models

The land-use approach to trip distribution, from housing and employment, leads first to journeys by all modes, rather than by a specific mode. Mode split and distribution are at the core of transport demand modelling. Both can be seen as choices, of mode or destination. In certain cases, including that of commuting in Wellington, they can be modelled jointly by a simple GLM. However, this is not appropriate when there are differences in the scale of uncertainty in the different choices. In practical models of mode choice, this is often represented by a nested logit model. A more general mixed logit model is being developed by researchers and it appeared that this could be calibrated by a HGLM, a development of GLMs.

1.1.8 Geospatial models

HGLMs can also represent spatial correlations in errors. K factors, often used to improve the fit of practical models, can be fitted as random variables, or correlations between them can be incorporated in models for testing.

1.1.9 Aggregate modelling from traffic count data

Travel data identifying the production and attraction zones of individual trips, such as the WTSM household interview survey, is usually costly and disruptive to obtain. Aggregate counts of traffic on road links or public transport services are much cheaper and easier to obtain, but cannot be used to calibrate trip distribution models by conventional means, or by GLMs. The techniques of matrix estimation have been developed to use count data, but not for model calibration. However, MVESTM, the matrix estimation computer program in the Cube (formerly Trips) suite, is founded on the same maximum likelihood principles as GLMs. It can fit a cost parameter, and allows great flexibility in specifying data and model structure separately. With some manipulation, these enable MVESTM to calibrate a trip distribution model from count data. The owners, Citilabs, kindly provided the program with the source code for its shell and a description of the intercept file structure. The path flow estimator developed by Bell and Grosso (1998) was also considered for this role.

1.2 Transport modelling

Transport modelling relates travel to land use, ie the type and density of houses, workplaces, shops and other development (Ortuzar and Willumsen 1994). It often considers travel by different modes, in particular private (car) and public transport (bus and rail). There are four stages in a conventional transport model.

- **Generation** calculates the amount of travel demand in an area from its land uses. The home ends of trips are known as productions, and the non-home ends (work, shopping) are known as attractions.
- **Distribution** takes the productions and attractions for each area and links the trip ends together, depending on the costs of travel between them.
- **Mode split** allocates trips between different modes according to their cost and travellers' access to them, eg car ownership.
- **Assignment** finds the routes between every origin and destination, and allocates the trips to them, giving the volume of traffic on each link and service.

The interaction of distribution with other stages is discussed in the literature review, section 2.5.

Traffic models are confined to vehicles and their assignment to the road network.

1.2.1 Zoning

Study areas are divided into zones. All travel to or from a zone is assumed to start or finish at a single point, the centroid, although there may be multiple hypothetical links (centroid connectors) between the centroid and different points on the representation of the real transport network.

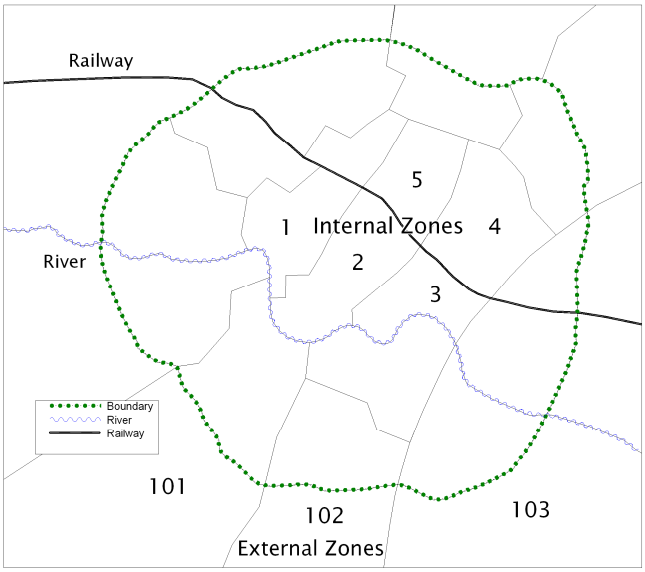
Zones are chosen to have homogenous land use, and are usually formed from an aggregation of administrative units, in particular census meshblocks, to aid the assembly of land-use data.

Zones are chosen so that all the traffic generated in them loads onto the same part of the network. For strategic transport models, zones can be formed around key junctions, but for traffic models and transport models where junctions are modelled in detail, zones are better formed and loaded onto the network between junctions.

Trips which start and end in the same zone are known as intrazonal and do not appear in the model of the real network. Larger zones will have more intrazonal trips.

Internal zones cover the whole of the study area and all travel between these zones is modelled. Outside the study area boundary, there are external zones to represent the origins and destinations of trips into, out of, or through the study area. Not all the trips between external zones are modelled.

Figure 1.1 Zoning example



1.2.2 Matrices

Patterns of travel are held in matrices. The rows and columns represent production and attraction zones, and the cells hold the numbers of trips between each pairing of production and attraction. The example in figure 1.2 shows 29 trips that are produced in zone 2 and attracted to zone 4.

Figure 1.2 Trip matrix example

Zones			Attraction										
			Internal							External			
			1	2	3	4	5	…	…	101	102	103	…
Production	Internal	1											
		2				29							
		3											
		4											
		5											
		:											
		:											
	External	101											
		102											
		103								Some			
		:								through			

Intrazonal

Travel is usually segmented by car availability and by purpose. Different household types produce trips for different purposes, depending on whether there are children to go to school or employees to go to work, and there are different patterns of attractors for different purposes – employment in one zone, shops in another.

Trip matrices can represent different household types, purposes and modes of travel.

1.2.3 Networks

Networks are represented by links which join at nodes. In highway networks, travel times are related to the volumes of traffic.

Public transport is represented by services running along the links. The path-finding process takes into account service frequency and interchange.

1.2.4 Costs

Costs of travel are calculated from the network, and affect choice of route, mode split and trip distribution. These are usually determined by a combination of distance, time and out-of-pocket costs such as fares or parking charges. This combination is known as the generalised cost. It may be scaled in terms of time, maintaining the same proportions between its components, and be referred to as generalised time.

Costs of travel between zones are held in cost matrices.

Costs for intrazonal movements are not modelled readily because they are not assigned to the parts of the real network represented in the model.

1.2.5 Directions and times of day

Productions and attractions refer to the home and non-home ends of trips, while origin and destination refer to the start and end of the trip. Trip generation and distribution are usually modelled over whole days, in terms of production and attraction. The resulting all-day matrices are factored into period matrices, typically representing:

- AM peak
- interpeak (IP)
- PM peak.

In the factoring, the directions are changed from production-attraction (PA) to origin-destination (OD). The majority of commuter trips have home as origin and work as destination in the AM peak; the pattern is reversed in the PM peak.

Person trips are also factored to vehicle trips using vehicle occupancy rates.

1.2.6 Surveys

Trip patterns showing the productions and attractions of trips are not readily observed.

There are two main ways of surveying travel patterns:

Household surveys approach people in their homes, and ask them to complete travel diaries of all the trips they make. They allow trip generation rates to be related to household characteristics, and cover all trips by all modes made from that household. However, they can be subject to under-reporting, particularly of short trips or optional travel. Personal attention by surveyors can improve the quality, but is expensive; at least two visits to a household may be needed, one to explain the trip diary and one to

check and collect it. Consequently sampling rates tend to be low, although absolute numbers are more important in calibrating a synthetic travel model.

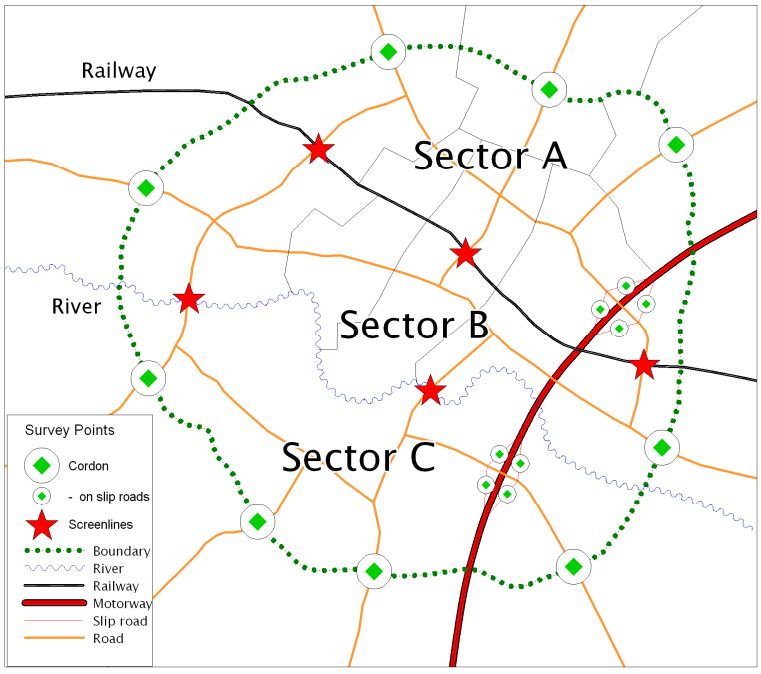
Census data is usually used to scale the sample up to the number of trips made by the whole population. The factoring process is sometimes called grossing up.

Roadside interviews stop drivers on the road, and ask them their origin and destination. The interview has to be brief, but usually covers the purpose at each end of the trip, which can identify the production and attraction related to land use. There is little opportunity to ask about personal or household characteristics.

Brief interviews by several surveyors working in one bay give a high sampling rate. All traffic, interviewed or not, is simultaneously counted to gross up the sample. Interview sites can be targeted to intercept certain movements, such as trips that are expected to use a proposed new road.

In modelling a whole study area, roadside interviews cannot intercept all trips, so they are conducted along complete cordons and screenlines. These separate the study area into sectors, so any trip between sectors has to pass through an interview site and may be sampled. Screenlines are laid out so that routes between sectors on either side of them pass through a screenline once and once only, and run along zone boundaries so sectors are made up of complete zones. Screenlines typically run along natural or artificial barriers, such as rivers, hill ranges or railway lines; they can become complex around motorways if interviewing is only allowed on the approach roads.

Figure 1.3 Cordon and screenline example



Study areas usually have at least a cordon of roadside interview sites on their boundary, to intercept trips by people living outside the study area, since these cannot be surveyed effectively by household interviews.

Movements wholly within the sectors bounded by cordons and screenlines should not cross any of them and are unobservable by the roadside interview survey (RSI); their cells in an observed trip matrix are null.

These need to be distinguished from movements between sectors, which must pass through an interview site and be sampled, but no trips have been recorded. This is an observation of zero.

On the other hand, some movements will pass through more than one screenline or cordon, and can be sampled and observed more than once. This multiple sampling has to be factored out in building the observed trip matrix from roadside interviews.

External trips into the study area must cross the cordon and may cross internal screenlines. External-to-external trips will cross the cordon twice, and possibly other screenlines, if they go through the study area. However, many external-to-external movements will not pass through the study area.

Figure 1.4 Partial matrix example – screenlines at which movements are observed

Sectors		Destination			
		A	B	C	External
Origin	A	None	Rail	Rail + river	Cordon + ?
	B	Rail	None	River	
	C	Rail + river	River	None	
	External	Cordon + ?			Some at cordon + ?

The resulting matrices will have blocks of unobservable cells within them. They are known as partial matrices. Trip distribution models can still be calibrated from them and can infill the unobserved blocks.

For multi-modal modelling, there may have to be complementary surveys of public transport usage, typically on-vehicle interviews or station surveys.

Counts of vehicles or passengers are cheaper than interviews, but do not give origins or destinations of trips, or relate their purpose to the land uses. Counts are often undertaken on sets of screenlines and cordons (even if interviews are not) to provide control totals for model validation. Ticketing information can give similar control totals for public transport.

Given a knowledge of routing from the assignment process, count data can be used to form or adjust matrices using methods known as matrix estimation.

Long-term count information, from automatic traffic counters, periodic counts or ticketing records, is used to adjust survey data for time-of-day, day-of-week and seasonal variations.

1.3 Trip distribution

Trip distribution is usually undertaken on trip productions (P) and attractions (A), relating to the home and non-home end of the trip. The terms origin (O) and destination (D) refer to the direction of travel in careful practice, but have a more general currency, including theoretical treatments of trip distribution. Trip distribution models are also known as destination choice or gravity models.

Gravity models can be applied in a wide range of spatial interactions relating two locations, such as:

- moving house
- changing jobs
- telephone calls
- marriage, relating the spouses' locations

and other forms of geographic analysis, but transport modelling is the major practical application.

1.3.1 Trip generation and balancing

Trip generation predicts the number of production P_i and attraction A_j trip ends in each zone.

In practice, the total estimated productions do not necessarily match the total attractions. Attractions are usually factored to match the total productions because production models, based on surveys of individual households and census data, are more accurate than attraction models, based on aggregate measures of broad planning data.

1.3.2 Model form

It is intuitively reasonable that the number of trips from zone i to zone j, T_{ij}, is proportional to the trip ends at the production P_i and attraction A_j.

$$T_{ij} \propto P_i \times A_j$$

Consistency as zones are split or amalgamated is difficult to maintain under any other formulation.

It is also reasonable for the tendency to travel to decrease with its cost. This is represented by a deterrence function, $f(C)$, so the form of the model becomes.

$$T_{ij} \propto P_i \times A_j \times f(C_{ij})$$

The cost, C, may be broadly defined as separation, which can include:

- the simple spatial separation of distance
- the state of the travel system reflected by actual travel time
- extra perceived deterrence of congestion, crowding or waiting for public transport
- out-of-pocket costs for fares, tolls or parking
- or socio-economic differences, as between a white-collar worker and a blue-collar job.

1.3.3 Deterrence functions

The original gravity model took its name from the inverse square law between distance and gravitational force. This is a particular case of the Power function

$$f(C) = C^{-\gamma}$$

Another form is the Exponential function

$$f(C) = e^{-\lambda C}$$

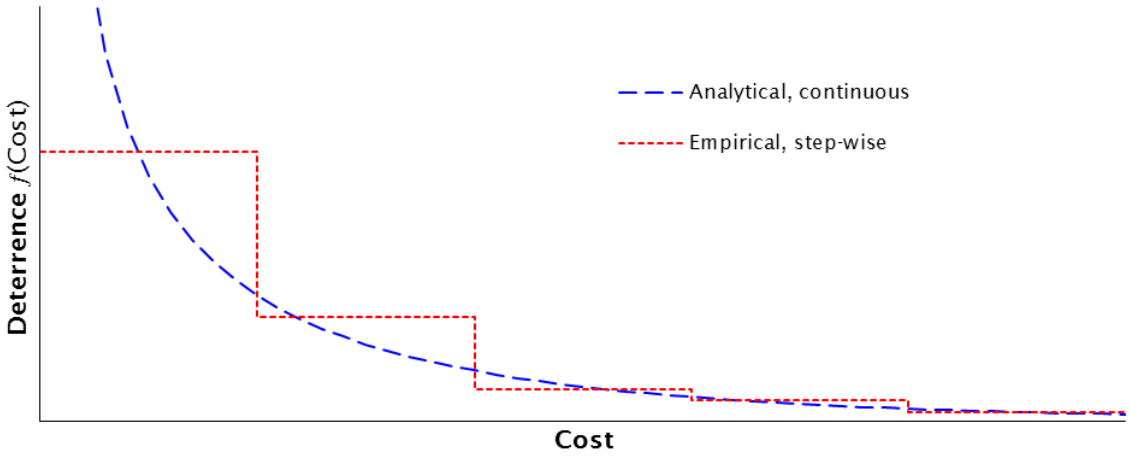
The two can be combined in the function

$$f(C) = C^{-\gamma} e^{-\lambda C}$$

This is a form of the gamma function, but is referred to as the Tanner function in this thesis to distinguish it from other uses of the term gamma.

As an alternative to these continuous analytical functions, an empirical step-wise function can be produced by splitting cost into bands, and calculating a deterrence factor for each band. This calibration needs only a relatively simple iterative balancing process.

Figure 1.5 Deterrence functions example



1.3.4 Balancing factors

A model of the form $T_{ij} \propto P_i \times A_j \times f(C_{ij})$ can be proportioned by a single factor, say K, giving

$$T_{ij} = K \times P_i \times A_j \times f(C_{ij})$$

so that the total number of distributed trips $\sum_{ij} T_{ij}$ will match the total number of trip generations, $\sum_i P_i$ or $\sum_j A_j$. However, this single factor cannot ensure that the trip distribution model will match the trip end totals for each individual zone, ie

$$\sum_j T_{ij} = P_i \quad \text{for all productions i, and}$$

$$\sum_i T_{ij} = A_j \quad \text{for all attractions j}$$

To meet these trip end constraints, there have to be two sets of balancing factors, one factor p_i for each production zone, and one a_j for each attraction zone. This gives

$$T_{ij} = p_i \times P_i \times a_j \times A_j \times f(C_{ij})$$

which is the usual form of the doubly constrained distribution model. This is used when there are constraints on the numbers of trips at both attractions and productions, typically for the journey to work where employees have to match the jobs available.

The singly constrained model has the form

$$T_{ij} = p_i \times P_i \times A_j \times f(C_{ij})$$

and is used when the attractions do not impose a constraint on the number of trips arriving, such as shopping. This distribution will not usually match predicted numbers of trip attractions for individual zones: $\sum_i T_{ij} \neq A_j$.

An unconstrained form of distribution arises from economic modelling and allows trip making to increase with the number and accessibility of attractions. It may be appropriate for occasional travel, such as long distance and leisure, where the constraint on overall trip-making implicit in the trip generation stage is less relevant.

1.3.5 Partial matrices

Partial matrices do not have observations in all their cells. These can arise from roadside interviews, where movements within the sectors bounded by interview screenlines and cordons are not interviewed. Distribution models can still be calibrated from partial data and the missing observations estimated or synthesised in the process.

Even 'full' matrices usually have incomplete observations of through movements, from external zone to external zone, because they may not be routed through the study area. External trips are often treated separately because the external trip generations, and in particular growth forecasts, cannot usually be derived from land-use predictions like internal zones.

Intrazonal trips are sometimes excluded while calibrating trip distribution to avoid the uncertainties in their costs.

1.3.6 Model fit and K factors

Synthesised trip distributions do not always give a good fit to observed data, with movements between whole sectors under- or over-predicted. This often happens with river crossings, where there appears to be a reluctance to cross a river to the other part of town which is greater than expected from modelled travel costs and the deterrence function. These mismatches can be adjusted by 'K' factors, which are applied to particular movements, such as those crossing the river. The model then becomes

$$T_{ij} = p_i \times P_i \times a_j \times A_j \times K \times f(C_{ij})$$

where, for instance, K may be 0.5 for movements that cross the river, and 1 otherwise.

K factors can improve the distribution model's fit to current observed data, but depend on the same empirical factors still applying under future forecast conditions. They are an abomination to theorists, but a very present help to practitioners.

K factors cannot apply just to whole rows or columns of a matrix, because their effects will be absorbed by the balancing factors, p_i or a_j. K factors can adjust distribution models to match total traffic counts along screenlines that divide sectors.

In the limit, K factors can be applied separately to every cell in a matrix (with one row and column of K values fixed to provide a reference level). This gives a marginal or incremental distribution model that can fit an observed matrix exactly, and uses the synthetic distribution model to factor the observed matrix in accordance with changes in trip ends and costs.

$$\text{Incremental forecast matrix} = \text{observed}_{\text{base}} \times \text{synthesised}_{\text{forecast}} / \text{synthesised}_{\text{base}}$$

This will retain the 'lumpiness' of a sparse observed matrix; in particular, all cells with zero observations will remain at zero. A major advantage of a synthesised matrix is a smoother distribution of trips between cells and hence around the network. A trip distribution cannot be calibrated simultaneously with such a set of K factors, because they can describe any matrix exactly and absorb any cost effects. Such K factors are calculated after calibration, as residuals.

1.3.7 Synthesis

There are two distinct elements in preparing a trip distribution model. Calibration finds the parameters of the deterrence function that best fits observations. Synthesis uses these to derive distributions for various scenarios, with different productions and attractions depending on different land-use developments, and different costs depending on different transport networks, with new roads or public transport services.

Synthesis can be a relatively simple process. An initial matrix is formed from the deterrence function of the costs, including any K factors. The trip generation model provides row and column totals of the trip ends. The rows and columns of the initial matrix are factored to match the trip end totals in a process that iterates between factoring rows and factoring columns. The process, known as Furness or Fratar balancing, is a common one in transport modelling, and the overall row and column factors are the balancing factors, p_i and a_j, in the trip distribution model.

1.3.8 Calibration

Suitable parameters for deterrence functions can be found by trial and error, synthesising distributions from observed trip ends and a variety of deterrence parameters, and comparing the results with the observed trip matrix. The prime criterion is the average trip cost and the distribution of trip costs is usually considered.

Empirical piecewise distributions can be fitted by extending the Furness balancing process into a third dimension, the cost bands. This can also be used as an intermediate step in fitting analytical functions.

For more complex deterrence functions with multiple parameters, such as the Tanner function or the fitting of K factors, trial-and-error methods become difficult. Advanced statistical methods can fit complex models in a single estimation process, and provide measures of the parameters' significance and the model's goodness of fit.

1.4 Wellington Transport Strategy Model

1.4.1 Development and structure

The Wellington Transport Strategy Model (WTSM) was substantially rebuilt for Wellington Regional Council by Sinclair Knight Merz and Beca Carter Hollings & Ferner Ltd from data from the national census and from travel surveys in 2001. The model is documented in a series of technical notes (TN) (Sinclair Knight Merz and Beca Carter Hollings & Ferner 2003).

The model is run in EMME/2, with trip end calculations in spreadsheets. Its 225 internal zones cover a diverse area with a population of 420,000. It extends from the centre of Wellington, New Zealand's capital, to the largely rural area of the Wairarapa which is connected to the rest of the study area only by a high road pass and railway tunnel through the Rimutaka Range.

The main survey was an interview of 2538 households. All residents over five years old (6953) filled in a self-completion travel diary for one weekday, covering 27,898 trips.

Trip distribution and mode split (summarised in table 1.1) are considered jointly, following practices evolved in the London Transportation Studies (MVA 1988). Their order was chosen so that cost parameters fitted a hierarchical choice model, with parameters increasing down the tree structure. The deterrence function for the distribution model is Exponential for consistency with choice modelling, taking the form

$$f(\text{cost}) = \exp(\text{constant} - \lambda \text{cost})$$

There are separate mode split/distribution models for:

- HBW – home-based work
- HBEd – home-based education
- HBSh – home-based shopping
- HBO – home-based other
- NHBO – non-home-based other
- EB – employer's business
- CV – commercial vehicle.

Trip productions are segmented by car availability:

- captive – no cars available in the household
- competition – more adults than cars in the household
- choice – cars for all adults

All production segments and modes are distributed between a single set of attractions, competing for them in doubly constrained distributions.

The following hierarchical geographic segmentation of trip movements is based on 16 sectors (nine urban, five rural) within six groupings of territorial local authorities (TLAs):

- intrazonal
- intrasector, urban
- intrasector, rural
- intra (TLA) group
- other (ie inter TLA group)
- CBD attraction.

Separate deterrence function parameters, both constant and cost 'parameter' λ, have been fitted for these geographic segments. Some combinations of geographic segments share the same parameters, but the combinations are not always the same for the constant and coefficient values.

Distribution and mode split are modelled between productions and attractions for a 24-hour weekday. Costs are taken from AM and IP networks in proportion to observed peak and off-peak travel for each mode, car availability and trip purpose.

Costs are derived from values set out by Transfund NZ, a forerunner of the NZTA. Values of time vary by purpose and car availability, reflecting public transport usage. Car costs comprise time, operating costs, parking charges and tolls, factored for vehicle occupancy. Public transport costs incorporate in-vehicle, walk, wait and interchange times plus fares.

There is simultaneous mode split and distribution for home-based work. For other purposes, motorised modes are split before distribution for each zonal production. The public transport modal constant can vary by TLA to match local mode shares and the coefficient of cost differs from that for car.

Slow modes (walk and cycle) are combined with a motorised mode for distribution, and later separated in a sub-mode split using an empirical diversion curve based on car travel distance. Treatment as a separate mode within a hierarchical choice structure gave inconsistent combined mode (logsum) costs.

Table 1.1 WTSM distribution segmentation and mode split hierarchy

Purpose	Production segments	Mode split	Sub mode factoring
HBW	Captive Competition Choice	*(Public transport only)* } Simultaneous	Public transport/car(pax)/slow Car/slow Public transport/slow
HBEd	Captive Competition+choice	*(Public transport only)* Before distribution	Public transport/slow Public transport/slow
HBSh	Captive Competition+choice	Before distribution Before distribution	} Car/slow
HBO	Captive Competition+choice	Before distribution Before distribution	} Car/slow
NHBO	Captive Competition+choice	Before distribution Before distribution	} Car/slow
EB	All	*(Car only)*	Car/slow
CV	*Separate model – not based on households*		

Since the formal modal split is between car and public transport, a binary logit model could be used for its calibration, rather than the more specialised multinomial logit.

Three time periods are modelled:

AM 7am–9am

IP 9am–4pm

PM 4pm–6pm

They are factored from the 24-hour matrices according to:

- purpose
- mode
- direction (to/from home)
- for some cases, movement within or between Wellington and other TLAs.

Vehicle occupancies are factored on a similar basis.

There is a peak spreading module, but as an incremental model, it does not affect calibration or synthesis of the base case.

The road network is loaded with an equilibrium assignment, with approach-based junction delays that reflect crossing traffic volumes. Link capacities are also applied.

The model loops through mode split, distribution and assignment. Costs are damped to help convergence.

Adjustment factors for each period matrix cell have been created by matrix estimation (MVESTM). These improve the fit of the model to link counts for more detailed studies in key areas. The WTSM is approved for use as a strategic model without the factors.

1.4.2 Running with EMME/2

Procedures for building and running the model are documented in a user manual, TN25.

The WTSM was set up to run under a size 2 licence for EMME/2. Canterbury University's licence is size 1, but all the prime dimensions of the WTSM fitted within that size. Some junctions were omitted from the turn table so it fitted within size 1; turning flows were not required for modelling these junctions. The assignment run thus under Canterbury University's licence gave a very close, but not exact, match with a previous assignment supplied with the model database. The changes to the turn table did not appear to affect the assignment.

The base year model, with turn bans omitted, ran in about six hours on a computer with a 300MHz CPU.

1.4.3 Household interview survey

GW supplied survey data as an Access database. The household interview survey (HIS) comprises three main related tables:

- households
- persons
- trips

with supplementary tables, notably vehicles. The database has good referential and logical integrity, and minimal missing data.

Although the database included all the data collected in the survey, it did not include coding to the definitions of mode, purpose and segmentation used to build the observed matrices for the calibration of the distribution model. It was only when GW supplied these observed matrices that all the definitions could be deduced and the matrices could be rebuilt exactly from the survey database. Uncertainty arose from features such as trip chaining which were suggested in early documentation of the study development, but not actually implemented.

1.4.4 Expansion factors

Expansion factors are attached to the HIS records. Those in the household and person tables appear to correspond to a two-stage expansion of surveyed households to census totals:

- first stratified by adults/employed
- then further stratified by children/TLA (TN9.1, section 3.4).

The factors in the trips table appear to incorporate these, plus adjustments for missing travel diaries calculated for persons stratified by age/employment (TN9.1 section 3.5).

Accumulating the expansion factors in the trips table replicates the observed matrices used to calibrate distributions.

1.5 Dataset studied

The original intention for this study was to take the form of the distribution model developed in the WTSM as an established base, and then fit alternative models available with GLMs, taking and testing one step at a time.

1.5.1 Commuting – home-based work (HBW) purpose

This study focused on commuting trips. This is the largest single purpose, providing the most data with which to examine distribution effects. It is relatively well defined, regular and well understood. Although travel for other purposes might be increasing more rapidly, commuting is still the core of the weekday morning and evening peaks, which set the regular demand for capacity in urban areas.

1.5.2 WTSM distribution/mode split calibration by GLM

In the WTSM, the distribution of HBW trips is combined with the modal split in a joint distribution/mode split (DMS) model. GW supplied the observed trip and cost matrices used in its calibration. Observed trip matrices could be rebuilt exactly from the household and roadside interview data. The totals of expanded trips for the observed trip matrices are shown in table 1.2.

Table 1.2 Observed home-based work trips

Household segment	Mode	WTSM observed matrices, as supplied	Compiled from interview data: household and roadside
Captive	Public transport	11,165.2	9839.71
	Slow		
	Car		1325.41
Competition	Public transport	25,618.2	25,618.25
	Slow	91,085.8	15,870.81
	Car		75,215.07
Choice	Public transport	18,274.8	18,274.95
	Slow		
	Car	110,732.6*	110,576.53

* Includes 155 external trips from household interviews

The WTSM form of distribution/mode split model was calibrated from the supplied trip and cost matrices by GLM. The fitted coefficients were very close to those quoted in TN17.1 and implemented in EMME/2.

Taking the costs from the DMS synthesis stage of the EMME/2 model gave an even closer calibration from the observed data, though still not exact.

Calibration by GLM on the synthesised trips from the DMS model stage with the costs from which they were synthesised recovered the DMS coefficients exactly, except for the constant for intrazonal public transport trips by competition households, of which there were only four observations.

Small changes in cost matrices are very hard to avoid during the development of a complex transport model, and are the likely cause of the small differences in calibration to observed data. It was concluded that GLMs replicated the calibration of trip distribution undertaken to build the WTSM, and reversed the synthesis of trip distribution which took place within it.

This WTSM DMS model applied 17 constants (K factors, plus base level) and 14 parameters (L factors) over a complex pattern of:

- three household car availability segments
- three modes, leaving a sub-mode split of the slow mode by empirical diversion curve
- five hierarchical geographic segments, plus CBD attractions.

Such a complex model was an unsuitable starting point for methodological development.

1.5.3 Mode and car availability

This study focuses on trips by car, which is the major mode. Costs for public transport trips become poorly defined for large rural zones where services are sparse, and these would also have been involved in a single all-mode distribution.

The three household segments were combined to give a single, simple matrix of HBW trips by car.

Having calibrated the full WTSM DMS model by GLM, the loss of fit in its reduction to a single HBW trip car distribution could have been examined. However, a thorough step-wise approach to statistical testing the decomposition and amalgamation presents a multiplicity of comparisons, many relating more to mode choice than to trip distribution. Although interesting, this would have leant more towards a review of the WTSM model than to developing the capabilities of GLMs, which was the objective of the study.

Car availability is fitted as a household characteristic in disaggregate analysis. Fitting mixed logit models by HGLM would have allowed a combined analysis of mode split and distribution. The WTSM form of hierarchical geographic segmentation is considered in the analysis of deterrence functions, section 4.4.

1.5.4 External trips from roadside interviews

The WTSM DMS model included external trips surveyed by roadside interviews on SH1 and SH2 near the study area boundary. The sampling rates at these surveys were much greater than those for the household interviews, from which the internal trips were derived. When different weights were applied to reflect this, there were marked changes in the fitted coefficient of a simple Exponential trip distribution model. This indicated a systematic difference between the fit of the model to internal and external trips.

1.5.4.1 Location of external generators

In the WTSM network, the centroid connectors from the study area boundaries to the single external zones are coded with a nominal length of 5km; the speed is fixed at 40km/h in common with all highway connectors. In reality, the next major centres are Levin, 20km beyond the study area boundary on SH1, and Palmerston North, over 60km beyond the boundary on both SH1 and SH2. The roads are mainly rural, with a 100km/h speed limit.

Under an Exponential cost deterrence function, changing the centroid connector cost does not change the relative attractiveness of different movements to and from a zone. However, if the Exponential function is not a complete model, as seems to be the case given the improvement in fit with the Tanner function, the distance beyond the study area boundaries to external generators will affect the distribution.

1.5.4.2 Location of surveys

Both roadside interview sites were located some distance inside the study area boundary, closer to the internal boundaries of the first internal zones within the study area. No adjustment for this was apparent in the survey data processing, apart from an edit check that rejected two interviews for not being internal-external movements.

The differences in fit that emerged from different weighting schemes appeared most marked in zones close to the study area boundary. Any difficulties with model convergence, whether in simple synthesis by Furness or complex fitting by matrix estimation methods, tended to appear in those zones remote from the centre of Wellington.

Although the roadside surveys intercepted relatively few movements or trips, their higher sampling rates could exaggerate their influence. The interviews did not collect household or personal data, so these could

not be included in disaggregate analysis together with the data from household interviews. The household interviews also offered a consistent (if smaller) sample of trips by public transport and slow modes.

For these reasons, external trips were omitted from the main analyses of this study, which were based solely on the HIS.

1.5.5 Core dataset

The core dataset for this study is the internal commuting (HBW) trips by car, recorded in household interviews conducted in 2001. Volumes are expressed as the number of person trips made during the whole 24 hours of a weekday.

They are generally formulated as a trip matrix defined by the production (home) and attraction (work) zones of the trips, as is conventional in the calibration and synthesis of trip distribution models.

Some documents referenced in this thesis use the terminology of origins and destinations, but this does not generally affect the theoretical approaches. The formal distinction between productions and attractions relating to the activity at trip ends, and origins and destinations, relating to the direction of travel, is addressed in chapter 8, where traffic counts by direction (and period) are used to calibrate a trip distribution.

The same dataset of observed trips is disaggregated from production zones to households, persons and trips in chapter 5, and aggregated to segments for computational purpose in chapter 7 and as a step on the way to calibration from traffic counts in chapter 8. In this thesis, the terms aggregate and disaggregate are relative to the zonal trip matrices which are the usual basis for the calibration and synthesis of gravity models, unless specified otherwise.

The dataset was re-formulated as pairings of home and workplace. This had little effect on the fitted coefficients, but a major effect on the statistics of fit. Weightings derived from the home-workplace formulation were therefore applied to trip-based data throughout this study. The home-workplace pairings were a step beyond tour-based modelling; trip chaining was considered for the WTSM, but not adopted.

The household interview survey sampled all internal trips, including intrazonal ones, so the trip matrix was fully observed and did not require partial matrix methods, although these can also be fitted by GLMs.

1.5.6 Costs

The costs used for calibrating deterrence functions in this research are those calculated for the synthesis of HBW car trip distributions in the WTSM base model. They are a combination of time and distance from the AM and IP assignments, plus parking charges for CBD attractions, expressed as generalised time in minutes. They were abstracted from EMME/2 after the final iteration of a damped distribution-assignment loop. Further details are given in appendix A. The contributions of individual components of the cost, ie time and distance, and AM and IP, are analysed in section 4.10.4.

The WTSM included parking charges in the costs for attraction zones in the CBD. This was more for the benefit of a mode split than trip distribution – under an Exponential deterrence function the distribution effect of such end charges is neutral. These parking costs were left in while there was the possibility of introducing other modes using WTSM formulations, up to the consideration of intrazonal costs in section 4.11.

They were removed to simplify the formulation and interpretation of alternative intrazonal costs when testing the sensitivity of deterrence functions to them.

In spatial analysis, costs were used as a measure of spatial separation between movements, as well as a deterrence to travel. Parking charges would have exaggerated separations in the CBD and were omitted.

The same costs were used for both separation and deterrence for simplicity and in the hope of allowing some reduction in the analytical problem. Costs representing separation were averaged by direction to provide symmetry in correlation matrices.

The same costs for deterrence, omitting parking charges, were adopted for aggregation/matrix estimation in the hope of recognising complementary results with spatial analysis.

1.6 Conventions adopted in this thesis

So far as is practicable…

The terms destination choice, trip distribution and gravity model are used interchangeably.

The usual effect of cost is to deter travel, or reduce the probability of choosing a more distant destination. An explicit negative sign is included in the cost term of deterrence functions

$$f(C) = \exp(-\lambda C) \text{ or } C^{-\gamma}$$

so that coefficients of cost or its logarithm are positive in this usual case. Larger coefficients imply a greater sensitivity to cost. The coefficient fitted by GLM will usually include a negative sign. The convention in other packages and documentation needs to be considered; it can differ between the Power and Exponential components of a Tanner function.

Groupings of zones are termed ‘sectors’ – a part of the study area. Groupings of PA or OD movements are termed ‘segments’ – a part of the matrix. The geographic segmentation applied in the WTSM is in accord with this usage.

K factors are constants in the deterrence function that vary between segments. The term L factor is adopted in this study for cost coefficients that vary between segments, by reference to K factors and to the use of λ for the coefficient of cost or sensitivity in many works on choice modelling. (The term M factor was contemplated for the ratio between cost coefficients at different levels of hierarchical choice in mixed logit modelling.)

2 Literature review

This review focused on the theory of trip distribution, properties of its components, and methods of calibration that have been published. While most practical models are documented to some extent, the material is not so widely disseminated or readily available, and tends to report the final form of the model rather than the methodology and analysis that led to it.

The main review was undertaken in the latter part of 2003. Some material which came to light up to 2010 has been added, but the review has not been revised comprehensively with the benefit of the subsequent improvements in electronic indexing and citation databases.

In this chapter, distribution is described as being between origins and destinations, rather than productions and attractions. This follows the nomenclature of the key theoretical treatments, without affecting their basis.

2.1 Theoretical bases

2.1.1 Entropy maximisation

> *You should call it entropy. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate, you will always have the advantage.*
> (Discussion between Shannon, Von Neumann and Tribus, cited in Tribus and McIrvine 1971)

The maximum entropy trip distribution is the OD trip pattern that occurs most frequently from all permutations of individual trips, given constraints of trip ends (and hence total trip numbers) and cost. It was developed by Wilson (1969), who applied it to other geographic and transport processes.

If a total of T trips are distributed with T_{ij} in each matrix cell, any of the T trips can be chosen to be the first trip in the first cell. The number of ways of choosing the next trip is T-1, and the number of ways of choosing all the T_{11} trips to go into the first cell is:

$$T.(T-1).(T-2)\ldots(T-T_{11}+1) = T! \,/\, (T-T_{11})!$$

However, the ordering of the T_{11} trips in the cell does not affect the overall trip pattern, so the number of distinct choices is less by a factor of $T_{11}!$:

$$(T! \,/\, (T-T_{11})!) \,/\, T_{11}! = T! \,/\, ((T-T_{11})! \times T_{11}!)$$

There are then $T-T_{11}$ trips left to choose from, and so the number of distinct ways to choose the T_{12} trips for the next cell is:

$$(T-T_{11})! \,/\, ((T-T_{11}-T_{12})! \times T_{12}!)$$

The number of ways to choose trips for the first two cells is thus:

$$\text{number for first cell} \quad \times \quad \text{number for second cell}$$

$$T! \,/\, ((T-T_{11})! \times T_{11}!) \times (T-T_{11})! \,/\, ((T-T_{11}-T_{12})! \times T_{12}!)$$

$$= T! \,/\, ((T-T_{11}-T_{12})! \times T_{11}! \times T_{12}!)$$

The number of ways over all cells thus becomes:

$$\omega(T_{ij}) = T! \,/\, ((T-\Sigma T_{ij})! \times \Pi(T_{ij}!)) = T! \,/\, \Pi(T_{ij}!)$$

and does not depend on the order in which the cells are filled.

For mathematical convenience, the logarithm of $\omega(T_{ij})$ is maximised.

$$\begin{aligned} \text{maximise } \omega(T_{ij}) &\Rightarrow \text{maximise } \log \omega(T_{ij}) \\ &\Rightarrow \text{maximise } \log T! \,/\, \Pi(T_{ij}!) \\ &\Rightarrow \text{maximise } \log T! - \Sigma_{ij} \log T_{ij}! \\ &\Rightarrow \text{maximise } - \Sigma_{ij} \log T_{ij}! \end{aligned}$$

since logT! is a constant because the total trips T are fixed by the trip end constraints.

This function $\omega(T_{ij})$ is maximised under constraints on the trip end totals

$$O_i = \Sigma_j T_{ij}$$

$$D_j = \Sigma_i T_{ij}$$

and on the total cost for the network

$$C = \Sigma_{ij} c_{ij} T_{ij}$$

where c_{ij} is the cost from i to j.

These constraints are included in the minimisation with Lagrangian multipliers α_i, β_j, and λ.

$$\text{Maximise } - \Sigma_{ij} \log T_{ij}! + \Sigma_i \alpha_i (O_i - \Sigma_j T_{ij}) + \Sigma_j \beta_j (D_j - \Sigma_i T_{ij}) + \lambda (C - \Sigma_{ij} c_{ij} T_{ij})$$

Differentiating to find the stationary point gives

$$\partial/\partial T_{ij} = - \partial(\log T_{ij}!)/\partial T_{ij} - \alpha_i - \beta_j - \lambda c_{ij} = 0$$

$$\partial/\partial \alpha_i = O_i - \Sigma_j T_{ij} = 0$$

$$\partial/\partial \beta_j = D_j - \Sigma_i T_{ij} = 0$$

$$\partial/\partial \lambda = C - \Sigma_{ij} c_{ij} T_{ij} = 0$$

One form of Stirling's approximation is

$$\log T! \approx T \log T - T$$

and its differential is used to give

$$\partial(\log T!)/\partial T \approx \log T + T/T - 1 = \log T$$

for substitution in the minimisation with respect to T_{ij} giving

$$- \log T_{ij} - \alpha_i - \beta_j - \lambda c_{ij} = 0$$

$$T_{ij} = \exp(- \alpha_i - \beta_j - \lambda c_{ij})$$

Rewriting

$$a_i = \exp(-\alpha_i) \,/\, O_i$$

$$b_j = \exp(-\beta_j) \,/\, D_j$$

gives

$$T_{ij} = a_i O_i b_j D_j \exp(- \lambda c_{ij})$$

which is the form of the doubly constrained gravity model with Exponential deterrence function. The trip end balancing factors are related to the Lagrangian multipliers that include the trip end constraints in the maximisation. The equations for the multipliers show their interdependence

$$\exp(-\alpha_i) = O_i \,/\, \Sigma_j \exp(-\beta_j - \lambda\, c_{ij})$$

$$\exp(-\beta_j) = D_j \,/\, \Sigma_j \exp(-\alpha_i - \lambda\, c_{ij})$$

and hence iterative procedures are needed to fit them.

2.1.1.1 Entropy

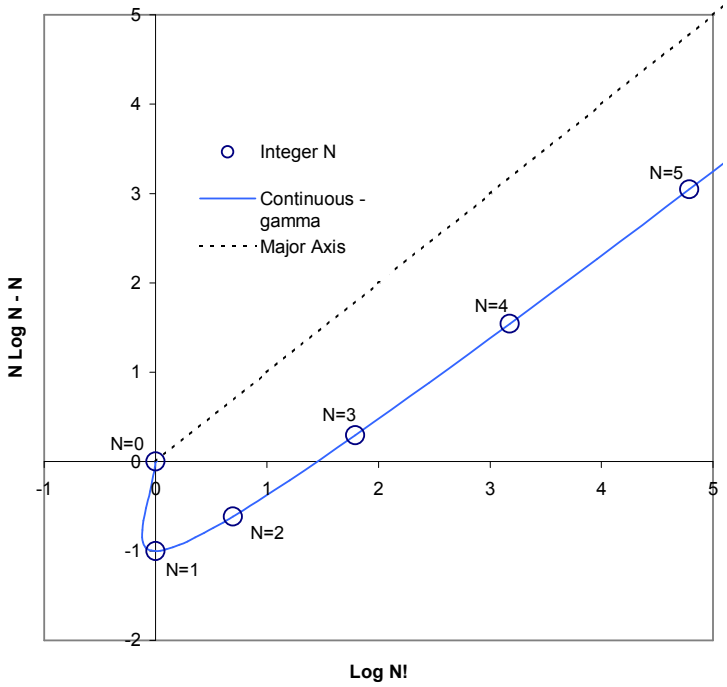
Wilson (1969) defined entropy as

$$h = -\Sigma_{ij}\, t_{ij} \log t_{ij}$$

where $t_{ij} = T_{ij} \,/\, T$, the proportion of total trips in each cell.

Maximising entropy is equivalent to maximising $\omega(T_{ij})$, the number of ways of allocating trips to cells, *assuming* the form of Stirling's approximation used, which is better for large numbers (figure 2.1). More terms in Stirling's series improve the approximation for smaller numbers.

Figure 2.1 Stirling's approximation



A routine to enumerate all possible permutations of trips has generated sets of distributions with the same trip ends, but without a constraint on network cost. The trip totals are small (<=9) because of factorial effects. One case has been found where $\omega(T_{ij})$ and H rank the distributions differently. The differences do not occur at the maximum entropy.

The routine allocated trips within trip end constraints, producing a different formulation from Wilson's (1969) for the number of permutations:

$$\omega(T_{ij}) = \Pi_i(O_i!)\, \Pi_j(D_j!) \,/\, \Pi_{ij}(T_{ij}!)$$

Given the trip end constraints on O_i and D_j, this is proportional to Wilson's (1969) formulation, $T! \,/\, \Pi_{ij}\, (T_{ij}!)$.

Entropy is a concept with similar forms of function used in information theory and statistical mechanics, for example representing the properties of gases by the interaction of molecules. The applicability of these concepts to the behaviour of individual travellers has been debated (CN32).

Erlander and Stewart (1990) define entropy as

$$H = \quad -\Sigma_{ij} (T_{ij} (\log(T_{ij}) - 1)) - 1$$

If T_{ij} is scaled to be the proportion of total trips, $t_{ij} = T_{ij} / \Sigma_{ij}T_{ij}$, this reduces to

$$-\Sigma (t_{ij} (\log(t_{ij})), \qquad \text{the form given by Wilson (1969)}$$

With this scaling, Erlander and Stewart (1990) show that for a given pattern of trip ends o_i and d_j, the maximum entropy is

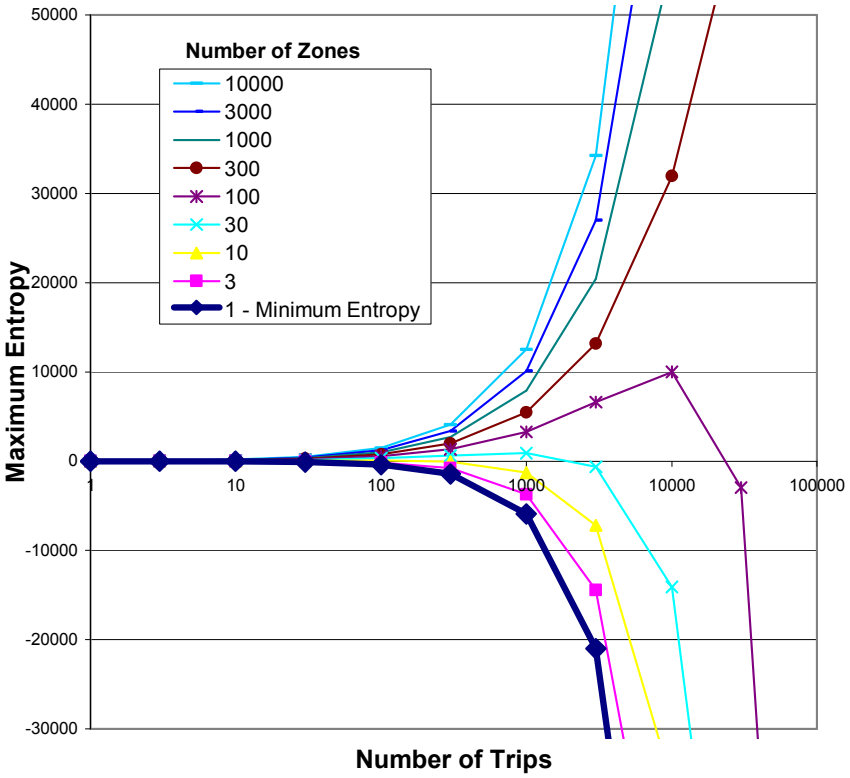
$$-\Sigma (o_i \log o_i) - \Sigma (d_j \log d_j)$$

and this occurs when $t_{ij} = o_i \times d_j$, ie all trips are spread out in proportion to the trip end totals.

This function itself has a maximum of log(I×J) when trip ends are spread equally among the I origin and J destination zones, ie $o_i = 1/I$ for all i, $d_j = 1/J$ for all j. The proportion of trips is then the same in every cell of the matrix.

Thus entropy is scaled by the number of cells in the matrix. If it is calculated from absolute trip numbers, rather than proportions, it is also scaled by the number of trips. Figure 2.2 shows how maximum entropy, with the same number of trips in every cell, varies with the number of zones and the total of trips.

Figure 2.2 Scaling of entropy



In contrast to these conditions for maxima, minimum entropy occurs when all trips are from one origin to one destination. It is zero when trips are scaled as a proportion of the total. Empty cells do not contribute

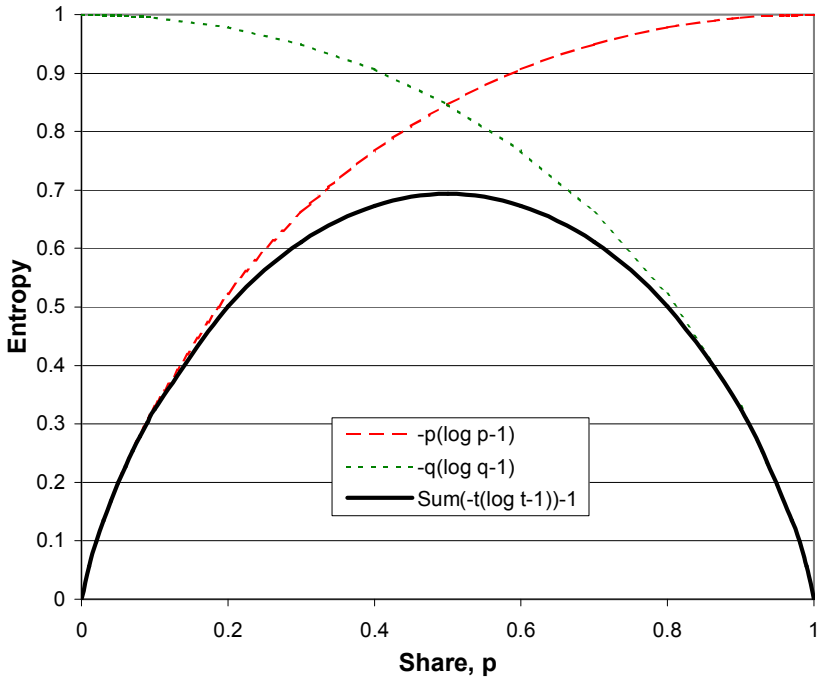
to the total entropy, by definition, so the minimum entropy for any number of zones, occurring with all but one cell empty, is the same as the entropy for a single zone. Thus in figure 2.2, the line showing maximum entropy for one zone is also the minimum entropy for any numbers of zones. This assumes all trip ends can be concentrated in one zone; there will be higher minimum for distributions constrained by other trip end patterns.

These limiting conditions are shown for a Wellington trip distribution in figures 2.4 and 2.5.

Both the upper and lower bounds of entropy vary non-linearly with the zoning system, the total number of trips and the pattern of trip ends. It is not a measure that is readily portable between models.

Maximising entropy produces a broad and even spread of trip ends between zones, and trips between cells. Figure 2.3 shows this for the simplest case, where trips are distributed between just two alternatives, with proportions p and q (p+q=1). The contributions of p and q to the sum of entropy are shown, together with the total entropy. This has its maximum at p=q=0.5, where trips are spread evenly.

Figure 2.3 Entropy in a two-way distribution



Returning to the original concept of the most frequent distribution of trips, this distribution of trips between two alternatives is equivalent to tossing a handful of coins to see the distribution between heads and tails. The most probable outcome is equal numbers, and the term $\omega(T_{ij}) = T! / \Pi(T_{ij}!)$ gives the relative frequency of P heads and Q tails when rewritten (P+Q)!/(P!Q!).

2.1.1.2 Cost constraint

While maximising entropy spreads trips across cells, minimising costs over the whole network tends to concentrate trips into a few low-cost cells. Intrazonals, which have the lowest costs, are used first for matching productions and attractions within each zone, and any imbalance is matched to the nearest zone(s) with a complementary internal balance. This is the classic transport problem in linear programming (LP), used for efficient distribution of goods from factories to customers, or moving spoil from cut to fill in earthworks.

However, these suppose the distribution of a single commodity by a single authority, whereas the distribution of person trips represents a myriad of decisions by individuals. The single commodity has the same value wherever it is produced or delivered, but individuals have a vast range of preferences about where they work, learn or shop. Thus there is an element of dispersion beyond the minimum cost distribution.

2.1.2 Optimisation

Erlander and Stewart (1990) considered distribution in the framework of a convex minimisation problem of the form

$$\text{Min } \Sigma_k x_k\{\log(x_k/x^0_k) - 1\} + f(y)$$

under a set of conditions on x and y, typically trip end totals.

The term $-\Sigma_k x_k\{\log(x_k/x^0_k) - 1\}$ can represent entropy, so the optimisation gives a maximum entropy distribution under a constraint on total network cost. Minimising network cost for a given entropy also results in a distribution model with an Exponential deterrence function.

This form of distribution is shown to give the best balance between maximum entropy (dispersion) and minimum cost. The cost coefficient λ represents the best trade-off between the two.

2.1.2.1 Relationship between entropy and cost in Wellington

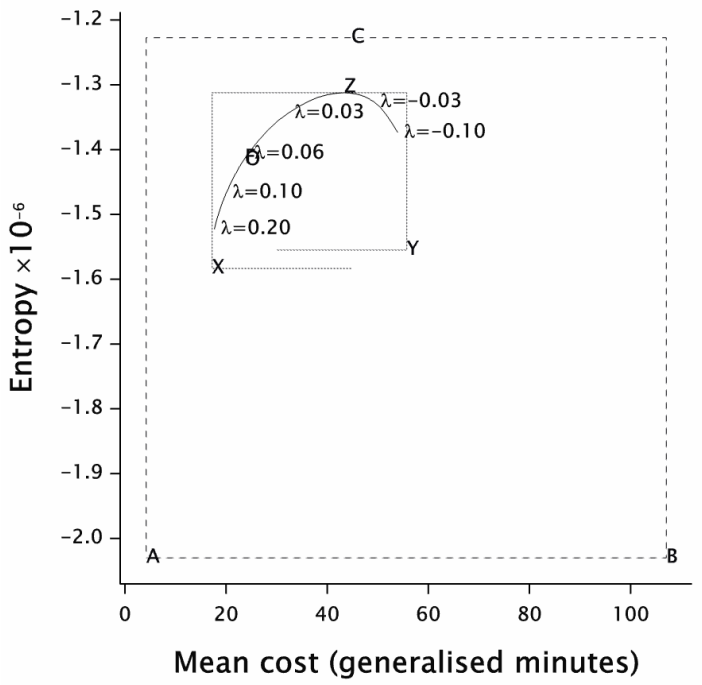
Figure 2.4 shows this relationship for a small set of Wellington travel data for HBW trips by car. Matrices synthesised for the base year have been compacted into nine zones, based on the 16 WTSM sectors, but excluding rural Kapiti Coast, Wairarapa and external sectors (13–16). The new zones and their trip ends are shown in table 2.1.

Table 2.1 Wellington distribution example – HBW private trips, 24-hour weekday

Zone	Location	WTSM sector	Productions	Attractions
1	Wellington south	1	37,696	23,665
2	Wellington west	2	18,856	13,925
3	Wellington centre	3	3294	51,488
4	Wellington north	4	21,170	10,170
5	Johnsonville	5	6023	3725
6	Porirua	6 & 7	19,941	11,094
7	Lower Hutt	10, 11 & 12	45,756	42,729
8	Kapiti Coast urban	8	12,392	10,088
9	Upper Hutt	9	17,504	15,747
Total trips			**182,631**	**182,631**

The costs are modelled highway costs weighted by (synthesised) trips. They are generalised in units of minutes and include parking charges.

Figure 2.4 Entropy and cost, Wellington



Both the outer and inner frames of the plot represent boundary conditions, including those discussed in section 2.1.1.1. The inner (dotted) frame is determined by the trip end distribution, whereas the outer (dashed) boundaries are set only by the overall scale of the distribution.

The left and right edges of the outer frame are simply the lowest and highest costs in the matrix, intrazonally within Johnsonville (4.21 minutes) and from Kapiti Coast to south Wellington (107.05 minutes) respectively.

The bottom of the outer frame is the minimum entropy for the total of 182,631 trips in the matrix. It occurs when all trips are from one origin to one destination, so the bottom left and right corners, A and B, represent all 182,631 trips being within Johnsonville, or from the Kapiti Coast to south Wellington.

The top of the outer frame is the maximum entropy for 182,631 trips in a nine-zone matrix, and occurs when the trips are distributed equally throughout all 81 cells. The average trip cost is then the simple average of all the cells in the cost matrix, 44.77 minutes, marked as C. These overall limits to entropy correspond to those shown in figure 2.2.

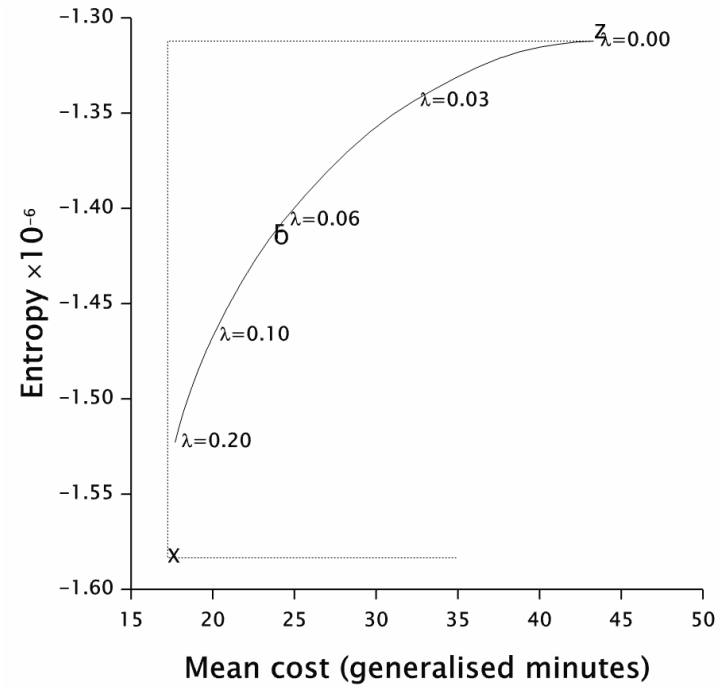
The inner (dotted) boundaries are set by the pattern of trip ends. The distance of the inner upper boundary below the outer one reflects the extent to which trip ends are not equally distributed between zones.

This lower maximum entropy is only achieved when trips are distributed in simple proportion to the trip ends. The average cost is then 43.33 minutes, at point Z on the plot.

Z is the turning point on the curve formed by Exponential distributions with varying cost coefficients, λ. At this point $\lambda = 0$, because λ is the gradient of the curve. It is also the Langrangian factor for the cost constraint, so with no weight given to costs, entropy can reach its maximum.

To the right of Z on the curve, λ is negative, so the curve maximises both entropy and cost. To the left of Z, entropy is maximised while cost is minimised; this is the usual region for transport models. It is shown, expanded within the inner limits set by trip ends in figure 2.5.

Figure 2.5 Entropy and cost, Wellington – detail



Other distributions are possible below the curve, but they will not be in the form of the gravity model with an Exponential deterrence function. The 'observed' trip distribution is marked at O, just below the curve. The distribution fitted from the observations, F, lies on the curve at $\lambda = 0.0637$, the fitted cost coefficient. Both have the same average trip cost, 23.73 minutes, because this is a constraint on the fitting process. The 'observed' distribution is very close to the curve because it is built up from synthesised distributions. These distributions have a complex segmentation of parameters, and have been amalgamated into just nine zones.

No distributions that meet the trip end constraints are possible above the curve.

The curve is plotted as far as distributions can be synthesised by GLMs. For large absolute values of λ, deterrence functions become very large or small, giving computational problems.

The left and right dotted inner boundaries are the minimum and maximum costs for any distribution given the trip ends. They are linear programming solutions to the classic 'transportation problem', discussed in section 2.1.1.2 'Cost constraint', and occur at X and Y on the plots.

2.1.2.2 Benefits

Erlander and Stewart (1990) showed that the trade-off between cost and entropy gives a way to value entropy and the dispersion (or unquantified utility) it represents. The net benefit, considering both this valuation of entropy and travel costs, is shown to be

$$-\left(\Sigma_i O_i \log O_i a_i + \Sigma_j D_j \log D_j b_j\right) / \lambda$$

Senior and Williams (1977) developed and used a similar measure, which they were able to apply in cases of restraint where the rule of half could not be applied.

2.1.2.3 Efficiency

Erlander and Stewart (1990) showed that several other criteria for optimisation, many related to those discussed above, lead to a distribution model with an Exponential deterrence function. In particular they show that this form of distribution is uniquely 'efficient': that is, under this distribution of cell probabilities p_{ij}, actual trip distributions T_{ij} with the same trip end constraints are always more probable if their network cost are less.

For any trip patterns T_{ij},T'_{ij} with the same constraints

$$\Sigma_{ij}\, T_{ij} = \Sigma_{ij}\, T'_{ij} = T,$$

$$\Sigma_{i}\, T_{ij} = \Sigma_{i}\, T'_{ij} = T\, \Sigma_{i}\, p_{ij},$$

$$\Sigma_{j}\, T_{ij} = \Sigma_{j}\, T'_{ij} = T\, \Sigma_{j}\, p_{ij},$$

if T_{ij} is more probable than T'_{ij} $\quad \Pi_{ij} p_{ij}^{T_{ij}} > \Pi_{ij} p_{ij}^{T'_{ij}}$

then T_{ij} is always less costly than T'_{ij} $\quad \Sigma_{ij}\, T_{ij} c_{ij} < \Sigma_{ij}\, T'_{ij} c_{ij}$ and vice versa

if and only if p_{ij} is distributed with an Exponential deterrence function, $\exp(-\lambda c_{ij})$.

2.1.3 Destination choice

Economic theories of choice, based on variations in values, have been extensively developed, notably by McFadden (1978). The theories find other transport applications in mode choice and stated preference methods, and in multi-route assignment, using either analytical (Dial) or stochastic (Burrell) methods. They were applied to trip distribution as a choice between destinations by Cochrane (1975).

Cochrane hypothesises that every attractor, such as a shop or workplace, has some utility. Travellers will choose the attractor with the maximum utility. The more attractors there are to choose from, the higher the maximum utility is likely to be.

Different travellers find different utility in each attractor; some favour one shop, some another. There is thus a spread in the perceived utility of each attractor, which may be represented as a probability distribution.

When considering the maximum amongst a large number (>>10) of such random utilities, it is only the distribution of the higher utilities, in the upper tail of the distribution, that is important. This is demonstrated at the end of this section.

For many common distributions, the upper tail approximates to a simple exponential function, where the cumulative probability $P(\text{Utility}>u) = \exp(-\lambda(u-m))$, where m is a measure of centrality and λ a measure of spread.

The probability of Utility<=u is thus $1- \exp(-\lambda(u-m))$ for a single distribution. The probability that every utility in n such distributions is less than or equal u is then

$$\Phi(u) = [1- \exp(-\lambda(u-m))]^n$$

This is the cumulative probability that the maximum utility from n independent distributions is equal to u.

$$\log \Phi(u) = n \log [1- \exp(-\lambda(u-m))]$$

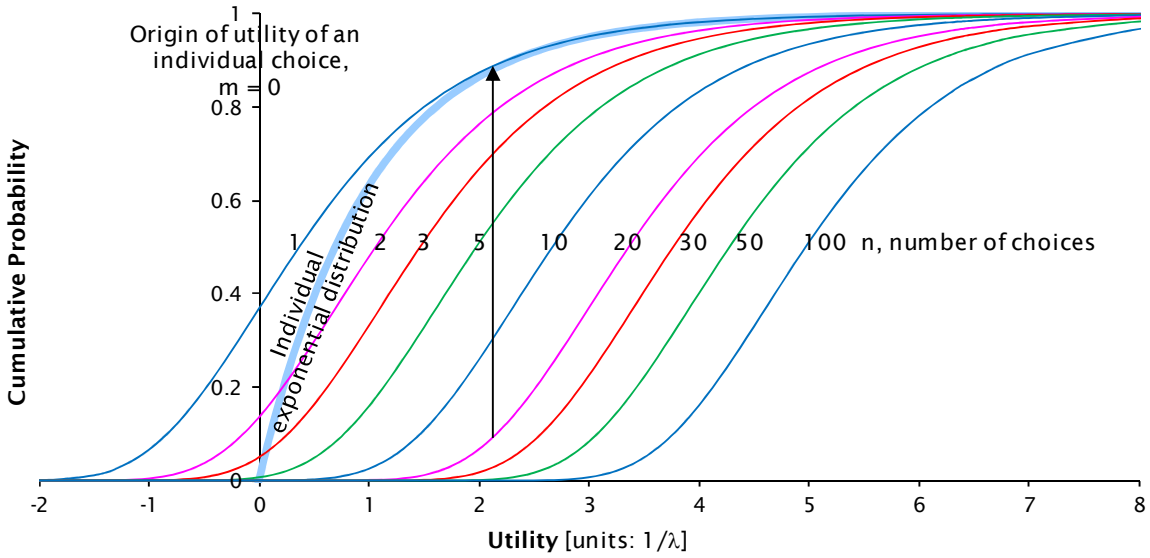
$$= -n [\exp(-\lambda(u-m)) + \tfrac{1}{2} \exp(-2\lambda(u-m)) + \ldots] \quad \text{expanding the series for } \log(1-x)$$

$$\approx -n \exp(-\lambda (u-m)) \quad \text{ignoring higher order terms since } \exp(-\lambda(u-m)) \text{ is small in the upper tail}$$

$$\Phi(u) = \exp [-n \exp(-\lambda(u-m))]$$

Figure 2.6 shows a family of such curves for the best utility from n choices.

Figure 2.6 Utility of choice



An individual exponential distribution is also shown. The 'best choice' curve for n = 1 fits well only in the upper tail, whereas it would be identical in the exact case. Curves for n<10 show a possibility of negative values, unlike the individual distribution.

However, for n = 20, the bottom 10th percentile is a utility of about 2. This must be the utility at almost the 90th percentile of one of the 20 individual distributions, as shown by the arrow in the graph. Thus the form of most of the n = 20 curve is determined by the upper tail of the individual distributions.

2.1.3.1 Gumbel distribution

Differentiation of the cumulative function, $\partial\Phi(u)/\partial u$, gives the probability density function

$$\varphi(u) = -\lambda n \exp[-\lambda(u-m) - n \exp(-\lambda(u-m))]$$

This is a Gumbel or type I extreme value distribution, shown in figure 2.7. These have

$$\text{mean} \qquad m + [\text{Log}(n) + 0.577]/\lambda$$

$$\text{and standard deviation} \qquad \pi/(\sqrt{6} \cdot \lambda)$$

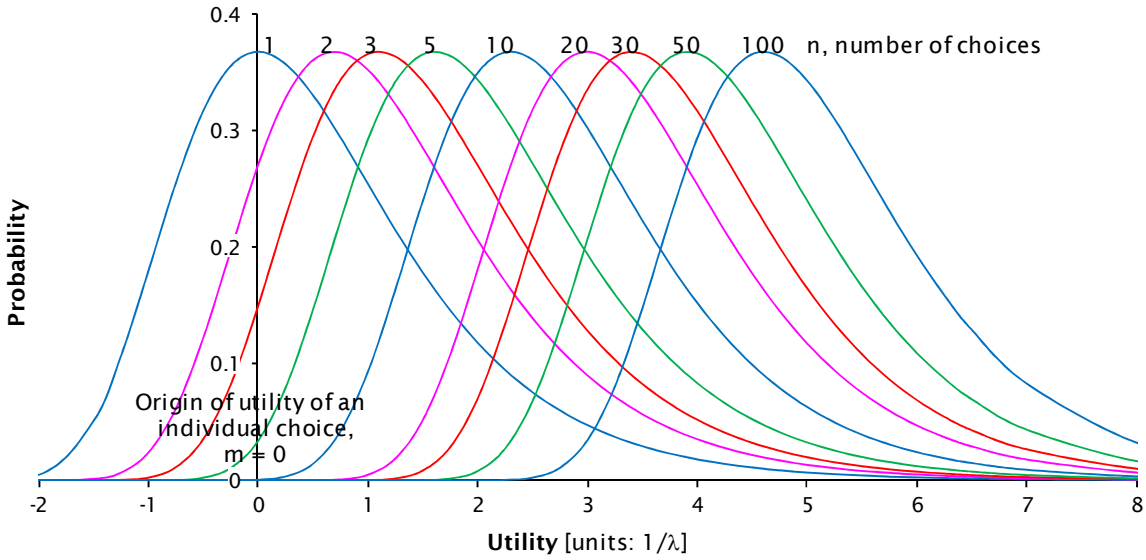
The position of the distribution thus depends upon the position of the original distributions m and the number of distributions n modified by the original spread in utility, λ. The spread of the distribution depends only on the original spread in the utilities λ.

The maximum of values taken from two or more Gumbel distributions is another Gumbel distribution: since the two source distributions can be seen as the maxima of n_1 and n_2 original individual distributions, their combined maximum will equivalent to the maximum of n_1+n_2 individual distributions –

$$m + [\text{Log}(n_1+n_2) + 0.577]/\lambda$$

This useful property applies *only* if both of the source distributions have the same spread, λ; the resulting distribution will have the same spread. The independence of the two source distributions is also critical.

Figure 2.7 Gumbel distributions



It may not be possible to identify and enumerate individual attractors in a zone, but this number can be taken to be proportional to other measures of zonal attraction A_j such as retail floorspace

$$n_j = hA_j$$

The net benefit or surplus s of travel to a zone is the utility of the attractors u less the cost of travel c.

$$s_{ij} = u_j - c_{ij}$$

Substituting these into Φ above, the cumulative distribution of the surplus for travel from zone i to zone j becomes

$$\Phi_{ij}(s) = \exp\left[-hA_j \exp(-\lambda(s-m+c_{ij}))\right]$$

with an equivalent form for the probability density function $\varphi_{ij}(s)$.

A traveller from i will go to j if that gives the most surplus. The probability can be derived from the probability that a surplus s can be found at j, and that the surplus at other zones k is less, $\Phi_{ik}(s)$, integrated over s.

$$P_j = \int \varphi_{ij}(s) \prod_k \Phi_{ik}(s)\, ds, \qquad \text{for } k \neq j$$

$$\text{or } P_j = \int \varphi_{ij}(s) \,/\, \Phi_{ij}(s) \prod_k \Phi_{ik}(s)\, ds \qquad \text{for k including j}$$

Integration and substitution gives

$$P_j = A_j \exp(-\lambda c_{ij}) \,/\, \Sigma_k A_k \exp(-\lambda c_{ik})$$

This is a common property of choice of maximum utility between several Gumbel distributions. The probability of choosing a destination is proportional to the exponential of its utility factored by the spread parameter λ. Including the origin trip end constraint

$$\Sigma_j T_{ij} = O_i$$

$$T_{ij} = O_i P_j$$

$$= O_i A_j \exp(-\lambda c_{ij}) \,/\, \Sigma_k A_k \exp(-\lambda c_{ik})$$

which is the form of the singly constrained gravity model. It is also the form of the logit model.

2.1.3.2 Double constraint

The extension to the doubly constrained model with constraints on destination trip ends

$$\Sigma_i T_{ij} = D_j$$

is akin to the Langrangian multiplier applied in the maximum entropy approach, but Cochrane (1975) offers an economic interpretation. A cost for each destination j, β_j, is hypothesised such that the destination constraints are met. The cost can be thought of as a congestion charge where there is competition for a small number of attractors (eg jobs) in an accessible zone, or a premium to attract people to an inaccessible one with many attractors. This assumes a market in which these costs can be traded to reach equilibrium. The cumulative distribution of the surplus then becomes

$$\Phi_{ij}(s) = \exp\,[-hA_j \exp(-\lambda(s - m + \beta_j + c_{ij}))]$$

and the probability of travelling to j is

$$P_j = A_j\exp(-\lambda(\beta_j + c_{ij})) \;/\; \Sigma_k A_k\exp(-\lambda(\beta_k + c_{ik}))$$

$$= A_j\exp(-\lambda\beta_j)\exp(-\lambda c_{ij}) \;/\; \Sigma_k A_k\exp(-\lambda\beta_k)\exp(-\lambda c_{ik})$$

and $$T_{ij} = O_i A_j\exp(-\lambda\beta_j)\exp(-\lambda c_{ij}) \;/\; \Sigma_k A_k\exp(-\lambda\beta_k)\exp(-\lambda c_{ik})$$

This is equivalent to the conventional form of doubly constrained distribution with balancing factors a_i and b_j

$$T_{ij} = O_i D_j a_i b_j \exp(-\lambda c_{ij})$$

if $b_j = KA_j\exp(-\lambda\beta_j)/D_j$ where K is the arbitrary scaling of the balancing factors. Then

$$a_i = 1 \;/\; K\Sigma_k A_k\exp(-\lambda\beta_k)\exp(-\lambda c_{ik})$$

$$= 1 \;/\; \Sigma_j KA_j\exp(-\lambda\beta_j)\exp(-\lambda c_{ij})$$

$$= 1 \;/\; \Sigma_j D_j b_j \exp(-\lambda c_{ij})$$ the interrelationship shown by Wilson (1969, equation 22)

Then the destination balancing cost $\beta_j = (\log K + \log(A_j/D_j) + \log b_j) / \lambda$

If the attraction measure A_j is well chosen, it will be proportional to the trip ends D_j, and A_j/D_j (eg workplaces/work journey) will be a constant. The term log K shows that the absolute costs are arbitrary and only relative costs between zones are of interest.

2.1.3.3 Consumer surplus

In the singly constrained case, the surplus from all trips can be obtained by integrating $\varphi_{ij}(s)$ and summing across all pairs of zones giving

$$S_T = \Sigma_i O_i[0.577 + \log(h\exp(-\lambda m)) + \log \Sigma_j A_j\exp(-\lambda c_{ij}))] / \lambda$$

This includes the arbitrary scaling of attractiveness h and utility of a single attractor m. These do not appear in the change in consumer surplus arising out of a change in trip costs

$$\Delta S_T = \Sigma_i O_i[\log \Sigma_j A_j\exp(-\lambda c^{after}_{ij}) - \log \Sigma_j A_j\exp(-\lambda c^{before}_{ij})] / \lambda$$

or for a doubly constrained distribution

$$\Delta S_T = \Sigma_i O_i[\log \Sigma_j A_j\exp(-\lambda(\beta_j + c_{ij})^{after}) - \log \Sigma_j A_j\exp(-\lambda(\beta_j + c_{ij})^{before})] / \lambda$$

or

$$\Delta S_T = \Sigma_i O_i[\log \Sigma_j D_j b^{after}_j\exp(-\lambda c^{after}_{ij}) - \log \Sigma_j D_j b^{before}_j\exp(-\lambda c^{before}_{ij})] / \lambda$$

substituting destination balancing factors $b_j = KA_j\exp(-\lambda\beta_j)/D_j$

2.1.3.4 Part constraint

The form of a simple unconstrained model is given by assuming that travel only occurs if there is a positive surplus. With a cumulative surplus to a single attractor

$$\Phi(s) = 1 - \exp(-\lambda(s-m+c))$$

the probability of travel is then

$$P(s>0) = 1 - \Phi_{ij}(0)$$

$$= \exp(\lambda m) / \exp(\lambda c)$$

which is a balance between the attractor's utility, m, and the cost of going there, c.

A more complex partly constrained model is given by applying the same condition of positive surplus on making a trip in the singly constrained case

$$P_j = \int \varphi_{ij}(s) \prod_k \Phi_{ik}(s)\, ds, \qquad \text{for } k \neq j \qquad \{\text{Cochrane 1975, equation 8}\}$$

The integral is then taken from 0 to ∞ instead of from -∞ to ∞, giving

$$P_j = A_j \exp(-\lambda c_{ij}) / \Sigma_k A_k \exp(-\lambda c_{ik}) \times (1 - \exp(-h \exp(\lambda m) \Sigma_k A_k \exp(-\lambda c_{ik})))$$

The first term gives a singly constrained distribution, as found previously. The second term is an elasticity of trip generation from zone i with costs of travel from the zone. It is a combined function of costs to all destination zones and applies to travel to all of them. Redistribution with cost changes to individual zones takes place within the first term.

2.2 Costs and deterrence functions

2.2.1 Exponential function

The major theories behind distribution models point to an Exponential deterrence function.

With the Exponential model, the proportional attractiveness depends only on the absolute difference in costs:

$$f(C_1)/f(C_2) = \exp(-\lambda C_1)/\exp(-\lambda C_2)$$

$$= \exp(-\lambda (C_1 - C_2))$$

Thus the differences between a $10 fare and an $11 fare, or between a $100 fare and a $101 fare, both give the same proportional deterrence effect. This makes the model insensitive to the base (origin) of the cost measurement: if $10 is added to every cost, by a consistent change of centroid connectors for example, the distribution remains the same. Changes in the scale of the cost measurement (about the same origin) can be compensated by rescaling λ.

The elasticity of the function $^{\Delta T/T} / _{\Delta C/C} = -\lambda C$, ie a function of the cost.

The attractiveness of the function continues to increase as trip costs decrease, but within finite bounds. The function curve is continuous through zero into the realm of negative costs, which might be encountered with composite costs or a utility with an arbitrary origin.

Sen and Smith (1995) show that by taking the deterrence to be a function of a broadly defined vector of separation measures, eg travel time, distance, out-of-pocket costs, or functions of them, the Exponential deterrence function applies to a wide range of distribution models. The main ones are discussed and shown in figure 2.8.

2.2.2 Power function

Another form of deterrence function is analogous to the physical process of gravitational attraction in having an inverse square law

$$\text{Trips} \propto 1/\text{distance}^2$$

This form is generalised in the Power deterrence function

$$f(C) = C^{-\gamma}$$

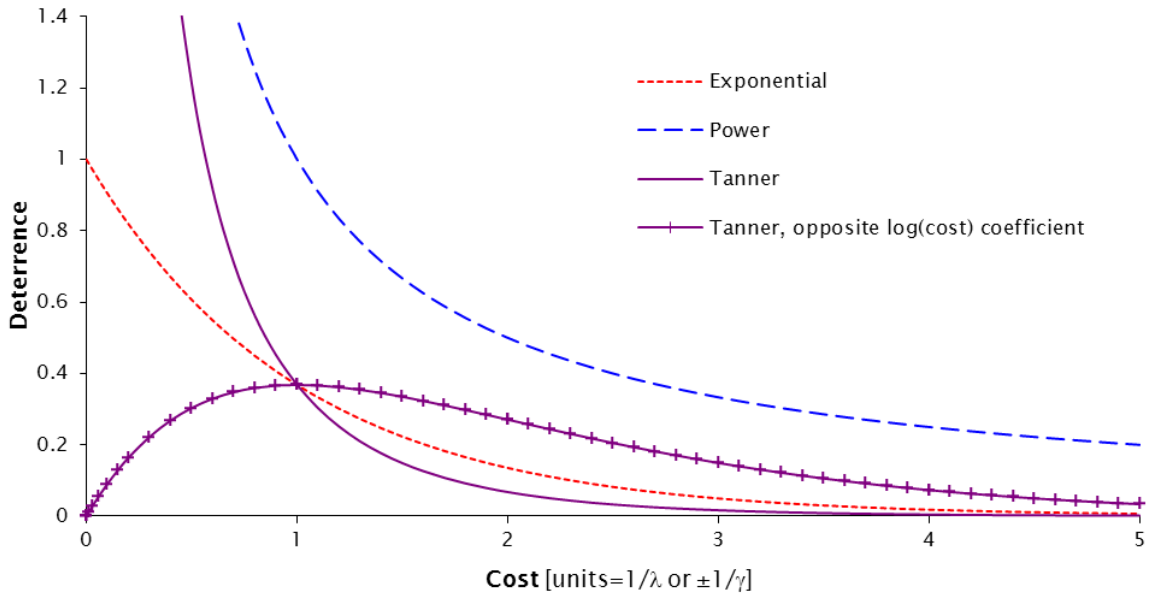
The exponent γ will be positive to give decreasing travel with increasing cost. Substituting the logarithm of the cost into an Exponential deterrence function can produce this form:

$$f(C) = \exp(-\gamma \log(C))$$

$$= C^{-\gamma}$$

Thus, while the Exponential model is constrained to reproduce the average cost of travel while maximising entropy, the Power model reproduces the average of the log(cost) of travel.

Figure 2.8 Deterrence functions



In the Exponential model, the proportioning of traffic depends on the difference in costs, C_1-C_2. The Power model thus depends on

$$\log(C_1) - \log(C_2) = \log(C_1/C_2)$$

ie the relative costs, to distribute trips. Thus the differences between a $10 fare and a $20 fare, or between a $100 fare and a $200 fare, both give the same deterrence effect. The model is insensitive to changes in the scaling of the cost measure, but is affected by changes in its origin.

The elasticity of the function $^{\Delta T/T} / _{\Delta C/C} = -\gamma$, ie a constant, which seems to be a favoured default assumption in economics.

The Power function tends to infinity as the trip cost diminishes to zero, unlike the Exponential. This can be awkward when dealing with short-distance trips like intrazonals, or separation measures that can take

negative values such as utility. The latter might arise in composite costs of several modes that are to be split after distribution.

Kirby (CN32) noted that for travel in a two-dimensional plane, a deterrence function of 1/distance is a better analogue of gravity's inverse square effects, 1/distance², in three-dimensional space. Tanner (1961; 1980) considered the implications of various gravity models for theoretical distributions of trip generators in a plane, eg point, grid, annular or continuous.

2.2.3 Tanner function

This function, introduced by Tanner (1961), combines the Exponential and Power functions

$$f(C) = C^{-\gamma} \exp(-\lambda C)$$

which is also a form of the gamma function. Again this can be re-written as an Exponential function

$$f(C) = \exp(-\lambda C - \gamma \log(C))$$

with a generalised cost that is a linear function of cost and log(cost).

The fitted distribution is constrained to reproduce the means of both cost and its logarithm. This helps reproduce the spread of trip costs as well as the mean.

The fitting might be thought of as ascribing importance between absolute and relative cost differences. However, in practice the exponent of the Power function γ can take a negative value contrary to this expectation. Alternatively, the function can be thought of as allowing non-linearity in the effects of cost, as in fitting a polynomial

$$a_0 + a_1 C + a_2 C^2 + \ldots + a_n C^n$$

Cost and its logarithm are naturally highly correlated, so there can be a high correlation between their fitted coefficients. However, this is typical of the components of generalised cost used in modelling (eg time, distance and fares). Sen and Smith (1995, p430) found the GLIM statistical package (for fitting GLMs, a subset of Genstat) to be relatively unaffected by multicollinearity between components of generalised cost.

When the exponent γ is negative, the function has a turning point and tends to zero as cost becomes small. This can help the fitting of short trips, but increasing trips with increasing costs does not rest easily with economic theory.

2.2.4 Empirical function – tri-proportional model

An empirical deterrence function can be generated by splitting the range of costs into bands and fitting a separate value for the deterrence of each band. If there are k = 1…K bands, the cost function can be represented by a set of dummy variables d_k which are 1 when the cost falls into band k, and 0 otherwise. With deterrence values a_k, the cost function becomes

$$f(\text{cost}) = a_1 d_1 + a_2 d_2 + \ldots + a_K d_K = a_k \qquad \text{where the cost falls in band k}$$

$$\text{or} \quad \exp(\lambda_1 d_1 + \lambda_2 d_2 + \ldots + \lambda_K d_K) = \exp(\lambda_k) \qquad \text{where } \lambda = \log(a)$$

Thus the function can again be regarded as a form of Exponential, with the dummy variables for the cost bands forming the linear components of generalised cost. The model is thus constrained to reproduce the mean value of each dummy variable. The dummies take values of 0 and 1, and their average is weighted by the number of trips, so this is equivalent to replicating the proportion of trips observed in each cost band.

This leads to the name tri-proportional, because the distribution now seeks to proportion trips to cost bands as well as to origins and to destinations. The cost-band deterrents play a similar role to the balancing factors for origins and destinations.

The distribution can be fitted by extending the iterative methods used for matrix balancing (Furness). These are less computationally demanding than the direct estimation of analytical functions, such as the Exponential or Power, and have been used as an intermediate stage. After an empirical function has been fitted iteratively to the mass of observations, the analytical curve is fitted through the resulting values, which only present one data point per cost band.

The empirical function need not be a monotonic decline in trip-making probability with cost that would be expected from economic theory.

The step-function between cost bands may generate noise and instability in a model as particular movements change between cost bands. Some software packages provide linear interpolation between the bands to mitigate this effect when synthesising trip distributions.

2.2.5 Other deterrence functions

Several other less common deterrence functions have been found. Most cannot be formulated as GLMs.

2.2.5.1 Box-Cox

The Box-Cox function

$$\exp(-\lambda(C^{\alpha}-1)/\alpha)$$

is provided for trip distribution synthesis in the VISUM software package (PTV 1997). It takes the form

$$\exp(-\lambda\log(C))$$

at the singularity at α=0, which reduces to the Power function $C^{-\lambda}$. At α = 1, it reduces to the Exponential function of cost, $\exp(\lambda)\exp(-\lambda C)$, so it can be seen as a mixture of the two functions, akin to the Tanner.

2.2.5.2 Box-Tukey

The Box-Tukey function is a generalisation of the Box-Cox function identified by Daly (2010) with an offset δ added to the cost C.

$$\exp(-\lambda((C+\delta)^{\alpha}-1)/\alpha)$$

2.2.5.3 Log-normal

The log normal function

$$\exp(-\lambda(\log(C+1))^2)$$

is provided in the OmniTrans (2006) software package for trip distribution synthesis.

2.2.5.4 Top log-normal

The top log-normal function

$$\exp(-\lambda(\log(C/\beta))^2)$$

is provided in the OmniTrans software package for trip distribution synthesis.

2.2.5.5 EVA1

The EVA model (from the German terms for production (Erzeugung), distribution (Verteilung) and mode choice (Aufteilung)) has been developed by Lohse at the Technical University of Dresden. It features joint mode split and distribution, and reconciliation of productions and attractions. Its implementation in the

VISUM software package offers the EVA1, EVA2 and Schiller functions as well as some of the more common functions described above.

The EVA1 function is

$$(\text{cost}+1)^{-\lambda/(\exp(\delta-\beta\text{Cost})+1)}$$

2.2.5.6 EVA2

The EVA2 function is

$$((\text{cost}/\beta)^{\alpha}+1)^{-\lambda}$$

2.2.5.7 Schiller

The Schiller function is the same as the EVA2, but with λ fixed at unity.

2.2.5.8 Lohse – scaling by minimum cost

VISUM offers a Lohse function

$$\exp(-\lambda(\{\text{cost}/\text{minimum}(\text{cost})\}-1)^2)$$

in its public transport model. This appears to be used for the choice of service, route or 'connection' in assignment, where a minimum cost path is usually well defined. In destination choice, the minimum cost will usually be for the intrazonal movement, which is heavily dependent on the zoning system and may thus be less suitable for scaling the costs to other destinations. Some effects of zone size might be absorbed into trip end balancing factors under doubly constrained distribution.

A more complex form, with λ as a function of the minimum cost, is also offered in VISUM.

2.2.5.9 Double Power – Tmodel

The TModel (1999) software package offered an extension of the Power function $1/C^{x}$ to the function

$$1/(C^{\alpha}+\delta C^{\beta})$$

with typical parameter values of $1<\alpha<3$, $-0.5>\beta>-4$, $50<\delta<1000$. With the second Power coefficient β negative, the probability of short trips is reduced, with a turning point at C=3.76 for α=2, β=−2 and δ=200. This allows for competition from walking over short distances in a traffic-only model. This function is included in VISUM as the 'TModel function' for synthesising trip distributions.

2.2.5.10 Double Exponential

A deterrence function of the form

$$\exp(-\alpha C)+\delta\exp(-\beta C)$$

was calibrated for the Christchurch Regional Model (1993), which was implemented in the TRACKS software package.

For heavy commercial vehicles, the fitted parameters were α = 0.208, β = 0.012 and δ = 0.0478 for costs in generalised minutes, which had components of 27.81 NZcents/minute and 52.80 NZcents/kilometre.

This is a form similar to the double Power. However, with all coefficients positive, there is no turning point or reduction in the probability of short trips. Instead, there is the concave form which is expected of cost damping.

2.2.6 Cost damping

Daly (2010) published a paper on 'cost damping' as part of a study to improve the UK Department of Transport's guidance on sensitivity parameters, given in unit 3.10.2 of its WebTAG website (www.dft.gov.uk/webtag/). Cost damping is a decrease in the sensitivity parameter or slope of deterrence

function with increasing cost or distance, compared with a simple logit model or Exponential deterrence function. The reverse effect of increasing sensitivity or slope is termed 'cost amplification'.

Daly found substantial empirical evidence for cost damping in transport studies in the UK and elsewhere, notably those carried out in Europe by RAND and Marc Goudry, and in studies of the value of time. The French Ministry of Transport requires Box-Cox tests, but non-linear functions are not accepted by US funding agencies, and the sensitivity parameters (or level of service coefficients) are tightly specified.

Daly reviewed microeconomic and random utility theories and did not find any strong reasoning for or against cost damping, or for particular forms. He proposed a practical kilometrage test, which requires that the total kilometrage driven decreases as monetary costs increase, and showed that this rejects the kinks in empirical functions, but accepts Box-Tukey functions with powers β in the range 0 to1. Since these limits represent the Power and Exponential deterrence functions, they are acceptable under the kilometrage test (although the Exponential function implies no cost damping, by definition). The test also accepts a linear mixture (with positive proportions) of acceptable functions, so the Tanner function (in its concave form) is also acceptable as a mixture of the Power and Exponential functions.

Daly found a range of functions had been used to model cost damping, with two principal classifications of the mechanisms:

- those that apply to the whole of a generalised cost, as opposed to those that apply to its individual components (eg time, distance, or monetary cost) or the differentials between them (eg value of time)
- those that are functions of 'fixed' variables such as distance or geographic segmentation, as opposed to those driven by 'policy' variables such as monetary cost.

He commended those acting on separate components of generalised cost, given the evidence of variation in the value of time, and those based on policy variables, to avoid the risk of obscuring policy effects.

Non-linear cost functions with varying sensitivity parameters may present practical problems for economic evaluations of consumer surplus, or applying the rule of half.

2.2.7 Intervening opportunities

The intervening opportunities model bases its deterrence function for a destination on the number of alternative destinations that are closer. In effect it uses a ranking of cost in place of cost itself. In doing so it loses information on the amount of difference in costs; it no longer knows whether closer destinations are much closer, or only marginally so.

Cochrane (1975, p41) showed that the intervening opportunities model could be equivalent to an ordinary distribution with Exponential deterrence function, but this depended on a particular relationship between the number of opportunities/attractors and the cost of travel. The Exponential distribution is one of a family of intervening opportunity models that are determined by the form of relationship between cost and number of opportunities. In practice, the actual planning data for the land use in each zone is incorporated in the model, so there is no need to assume analytical forms for density patterns.

Following suggestions by Wills (1986), Tamin and Willumsen (1988; 1989) used a Box-Cox transformation to determine whether the fitting of data favoured one of two intervening opportunity models or an Exponential cost deterrence model.

2.3 Trip ends and balancing factors

2.3.1 Accessibility

Kirby (1970) showed that balancing factors a_i and b_j are the mean of the inverse of the deterrence function, weighted by the fitted number of trips, across the trips to or from the zone to which the balancing factor applies.

$$a_i \propto \sum_j \{T_{ij} \,/\, f(C_{ij})\} \,/\, \sum_j T_{ij}$$

This is similar to the form of accessibility suggested by Hansen (1959)

$$\text{Accessibility}_i \propto \sum_j D_j \, g(C_{ij}) \,/\, \sum_j D_j$$

where $g(C_{ij})$ is a function of cost that measures accessibility. The inverse of deterrence $1/f(C_{ij})$ may be seen as a measure of accessibility.

However, the balancing factors are weighted by the trips distributed from the zone whose accessibility is being measured, and thus incorporate an element of traveller's choice and other factors affecting the distribution process. By comparison, Hansen's accessibility measures are weighted by all attractions of the destination zones. They are thus purely functions of the location of trip ends and the costs of travel between them; they do not reflect how travellers respond to these patterns in their trip distribution.

2.3.2 Marginal costs

Erlander and Stewart (1990, section 6.6) showed that with an Exponential trip distribution, excluding any value of entropy as estimated in section 2.1.2.2, the marginal network cost of adding one extra trip end from i and one extra trip end to j is

$$(\log O_i a_i + \log D_j b_j + H) \,/\, \lambda$$

This marginal cost is different from the cost of one trip from i to j, C_{ij}, as it allows for redistribution when the new trip ends are added at i and j. It assumes that all costs C_{ij} are constant, and are not affected by the marginal changes in traffic.

This is separable into origin costs, destination costs and general costs that are not specific to location. This suggests planning new residential development in zones with low origin costs, and employment or other attractors in zones with low destination costs. Erlander and Stewart offer an example in Sweden where this approach improves the balance of housing and employment.

However, economics suggest that competition will be causing equivalent adjustments to house prices and wage rates.

2.3.3 Shadow costs

The marginal costs are akin to the shadow costs in linear programming (LP) solutions. These are also separable costs, c_i and c_j, which are equal to the actual costs C_{ij} for the i to j movements that are part of the optimum solution: $c_i + c_j = C_{ij}$. For all other i to j movements, the shadow costs are less than the actual costs, $c_i + c_j < C_{ij}$, otherwise overall cost could be reduced further by including them in the solution. This is hardly surprising since trip distribution is cost minimisation like LP, under an additional constraint on entropy.

If actual travel costs C_{ij} are separable over parts of the cost matrix, there are alternative solutions to the LP problem, and computational difficulties arise.

2.3.4 Uncertainty in trip ends

Erlander and Stewart (1990, section 6.8) showed that trip end estimates did not have to be treated as absolute constraints. They could be considered as having a range of values, with upper and lower bounds. However, optimisation was then likely to set many totals to these bounds, and it was unsatisfactory for so many to be taking extreme values. An alternative, called the target value method, was to minimise the difference between the trip end totals and the target values as part of the optimisation.

The function of difference that is minimised is $\sum T\log(T/T_0)$, where T is the fitted trip ends and T_0 is the target value. The fit of origin and destination constraints, $\sum O\log(O/O_0)$ and $\sum D\log(D/D_0)$, can be given different weights in the overall optimisation. This offers a more flexible solution to the problem of forecast total productions not matching total attractions, $\sum O_0 \neq \sum D_0$; the conventional approach is to factor attractions to match productions.

2.4 Error components

2.4.1 Continuous and integer terms

Both entropy and choice theories are founded in integer arithmetic, with individual trips being allocated to matrix cells, or individual travellers choosing the destination that maximises their utility. With trips allocated at random, or individuals' tastes varying independently, this gives a Poisson type of process when the trips are counted (Sen and Smith 1995, section 3.6).

The resulting Exponential distribution model generates continuous values of expectations or probabilities. This is convenient in synthesised matrices since they smooth the distribution – small volumes can be split sensibly amongst any number of zones. No value would keep its integrity for long in a practical model full of factors and ratios, though it might add to the noise in iterative processes.

A continuous model is also more mathematically tractable. In LP terms, it is easier to search for the optimum along boundary conditions in continuous space without having to check each nearby integer. Exact integer distributions have been investigated by Holmberg and Joernsten (1989).

2.4.2 Large and small numbers

Most theories of distribution make some appeal to the large numbers of trips in a study area.

The allocation of a finite integer number of trips according to the continuous probabilities of distribution theory, considered above, will produce a 'population settlement' error between any actual population counts and theoretical expectations. With large populations such variability is small, and in practical terms unlikely to be the worst difference between distribution theory and actual behaviour.

However, while the population of trips may be large, the sample of them on which models are calibrated is usually much smaller. In the case of the WTSM, the trip sample was about 1/60 of weekday trips. Table 2.2 gives the raw counts of observed trips and the 'grossed up' values that estimate the whole population.

Table 2.2 Sampling of trips – home-based work, 24-hour weekday

Household segment	Mode	Grossing factor	Total		Per zone		Per cell	
			Count	Grossed	Count	Grossed	Count	Grossed
Captive	All	86.9	120	10,427	0.5	46	0.002	0.21
Competition	Car+slow	62.4	1279	79,799	5.7	355	0.025	1.58
Competition	PT	66.4	362	24,042	1.6	107	0.007	0.47
Choice	Car	59.6	1909	113,797	8.5	506	0.038	2.25
Choice	PT+slow	63.3	341	21,591	1.5	96	0.007	0.43
All	**All**	**62.2**	**4011**	**249,656**	**17.8**	**1110**	**0.079**	**4.93**

Source: Household interview survey for 225 internal zones

This sampling will contribute a substantial component of error. Because it arises from the counting of observed trips, it has a Poisson distribution. This applies on the scale of the raw count data, rather than data grossed up to represent the population trip matrix.

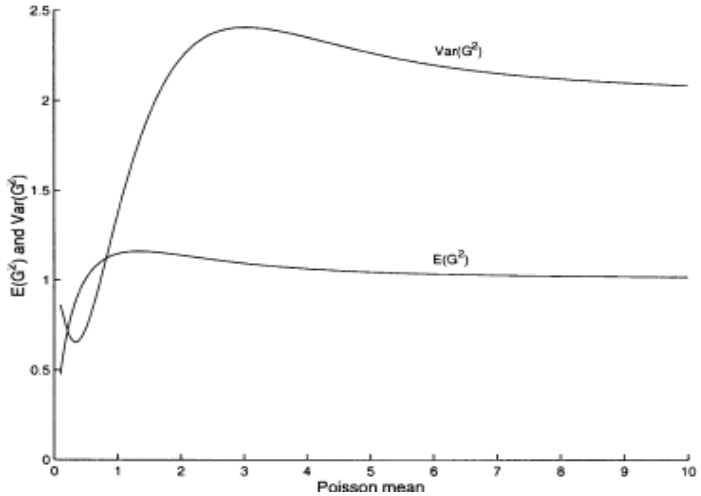
With a fairly consistent sampling factor in the WTSM, sampling and population settlement errors will have similar patterns on different scales. Consideration of the hypergeometric distribution (see section 2.4.8) suggests that population settlement effects compensate for sampling from a finite population, at least in the first order. With similar patterns, any residual effects of settlement error could appear as minor over- or underdispersion in the Poisson distribution of the sampling, and McCullagh and Nelder (1989, p199) state that modest overdispersion may be neglected.

In practical terms, it is unlikely that population settlement error can be distinguished from sampling error, which is by far the larger.

2.4.3 Sparseness

In modelling accidents which have a similar Poisson process, Wood (2002) noted that when there is less than one event per observation, statistics for testing the fit of GLMs no longer approximate well to known distributions. In particular, the mean deviance falls below its expected value of unity for a Poisson process, as shown in figure 2.9. To counter this, Woods suggests combining observations. For the counts of trips per cell in the HBW segments and modes given in table 2.2, his paper suggests cells need to be amalgamated into groups of 40 to 100.

Figure 2.9 Deviance (G^2) characteristics for sparse data



Source: Wood (2002, fig 5)

McCullagh and Nelder (1989) also recognise the problem, pointing out that the deviance of binary data, where all observations are 0 or 1, is utterly uninformative about the goodness of fit of the model (p120). However, they also note that even if the deviance of the final model does not approximate well to a known distribution, the change in deviance as variables are added to or dropped from the model may still provide a reliable test statistic (pp36, 119 and 122). Hence the statistical models are better used in comparisons between one another, rather than considering individual models. This is the same as good practice in transport modelling.

Sen and Stewart (1995, p361) note that maximum likelihood estimates can be derived from trip end totals and network cost, and that these sufficient statistics are relatively large numbers compared with individual cell counts. The form of Whittaker's formula (see section 2.4.10) is consistent with this view.

However, there are only 8.5 observed trip ends per zone for the HBW choice car segment (table 2.2), which is still not safely into the realm of large numbers, and the HBW captive segment has only 0.5 observations per zone, so many zones (at least 50%) will have no observations.

2.4.4 Grossing/factoring

It is the counts of observed trips that are Poisson distributed, rather than any values grossed up to represent the total population; however, the gravity model represents the total population. The two will be related by the grossing factor, F

$$\text{Population} = F \times \text{count}$$

Even if a survey is designed to give a constant sampling fraction, corrections for missing data and other practicalities usually produce variations in the final grossing factor across the matrix.

GLMs can be set up to represent a Poisson error in the observed counts while fitting a distribution model to the grossed-up population of trips in two ways (Sen and Smith 1995, section 5.9.3). This is done by modelling:

1 the grossed-up population of trips as the dependent (Y) variable and adjusting the errors by including a weight of 1/F

2 the observed counts as the dependent (Y) variable and factoring the linear function of the X variables by an offset of –log(F).

This assumes the grossing factors are known exactly, but they may be an additional source of error (CN07). Since grossing factors are usually the ratios of large well-observed numbers such as census totals or long-term traffic counts and totals of survey observations, the error in estimating these factors will usually be small compared with sampling error.

2.4.5 Grouping effects

In the WTSM, distribution is modelled over a 24-hour weekday. A simple journey pattern, with one trip from home to an attractor and a corresponding return trip home, is counted as two trips between the same production (home) and attraction zones. The two trips are converted back to origin-destination movements and allocated to periods within the day by factoring later in the model synthesis process.

The WTSM is built on person trips, rather than vehicle trips. Any passenger in a car is liable to generate multiple trip counts between the same production and attraction zones, the same as the driver. The pattern becomes complex where driver and passenger(s) have different purposes in different locations. Person trips are converted to vehicle trips using vehicle occupancy factors before highway assignment.

Both of these are normal practice in transportation modelling. At the distribution stage, they will group counts together more than would be expected in a pure Poisson process, giving an over-dispersed Poisson distribution.

2.4.6 Non-identifiability

Murchland (CN44) encountered issues of non-identifiability, where there is a lack of common information to link sets of zones. This applies particularly to partial matrices built from interviews along cordons and screenlines, where whole blocks of cells are unobservable. These should be distinguished from zero counts of movements that have to pass through the interview points and so are observable, but no such journeys were interviewed in the sample.

A simple case of a partial matrix with unobservable blocks is shown in table 2.3. It could arise from interviewing on a screenline between sectors A and B. There is no data linking the two blocks of surveyed data, so it is not possible to distinguish the relative attractiveness of the sectors as origins, or as destinations.

Table 2.3 Partially observed matrix

Origin\destination	Sector A	Sector B
Sector A	Unobservable	Surveyed trip data
Sector B	Surveyed trip data	Unobservable

It is a greater issue in fitting user-defined deterrence functions, where the costs are split into bands, and a separate parameter is fitted for the deterrence of each band. There then needs to be common information between cost bands to show their relative deterrence. These conditions do not arise in the WTSM, where a continuous deterrence function is fitted to a fully observable matrix.

In fitting GLMs, non-identifiability may appear as an issue of the rank of the data matrix input to the regression process.

Erlander and Stewart (1990, section 3.6) are careful in their analyses to distinguish cells that are forced to zero by structural constraints, but this does not appear to affect non-identifiability.

2.4.7 Multinomial distribution

The multinomial distribution occurs when there is a constraint on the total of a series of otherwise independent counts, such as the row or column total of a matrix. Sen and Smith (1990, section 5.2.4) show that, for estimating distribution model parameters, individual cell counts can still be regarded as Poisson processes. The two distributions give the same maximum likelihood estimators.

2.4.8 Hypergeometric distribution

The hypergeometric distribution arises from sampling a finite population without replacement.

The variance of an observed count of interest within the sample is

$$\text{Var}(np) = np(1-p)(N-n)/(N-1)$$

where

- p is proportion of trips of interest (eg for i to j movement, $T_{ij}/\sum T_{ij}$)
- n is sample size
- N is population size

The first term, np, is that for a simple Poisson process.

The second term, (1-p), allows for the finite sample within which the trips of interest are counted. For one cell in a matrix, it will usually be close to unity. With the first term, it gives the variance for a binomial or multinomial distribution, discussed above.

The remaining term reduces the variance as the unobserved population decreases. At the limit, there can be no variance in a 100% sample, because the whole population has been measured.

In sampling a population of distributed trips, this reduction for sampling a finite population is balanced by the 'population settlement' error, at least in the first order, as follows:

The sampling error in an estimate of the population of trips in a cell Np is

$$\begin{aligned}
\text{Var}(Np) &= \text{Var}(np) \times N^2/n^2 && \text{grossing up from the sampling variance} \\
&= np(1-p)(N-n)/(N-1) \times N^2/n^2 && \text{using the hypergeometric form above} \\
&\approx Np\,(N-n)/(N-1) \times N/n && \text{ignoring the multinomial constraint term } (1-p)
\end{aligned}$$

The 'population settlement' error as the whole population of N distributes itself, with probability p in the cell of interest, is Var(Np) = Np from the Poisson distribution.

Simply adding the two variances gives

$$\begin{aligned}
\text{Variance} &= Np\,[1 + (N-n)/(N-1) \times N/n] \\
&= Np\,[n/N + (N-n)/(N-1)] \times N/n \\
&\approx Np\,[n/(N-1) + (N-n)/(N-1)] \times N/n, && N>>1 \\
&= Np\,[N/(N-1)] \times N/n \\
&\approx Np \times N/n, && N>>1 \\
&= np \times N^2/n^2
\end{aligned}$$

which is the variance for a plain Poisson distribution of the sample, np, grossed up by N/n.

The hypergeometric distribution has been used in the European MYSTIC project (The Methodology and Evaluation Framework for Modelling Passenger and Freight on Transport Infrastructure Scenarios using the European Transport Policy Information System; IVT Heilbronn & Sandman Consultants Ltd 1999; Gaudry 1999), and in the ERICA software developed by Peter Davison for the project.

The software builds matrices from multiple sources; the same trip may be observed by different methods (household, interviews screenline, freight surveys) with different sampling fractions and possibly by different studies in different countries. The relative accuracies of the sources are used to give the best joint estimate.

At a practical level, most of the WTSM trip data is observed once only in the single household survey. The sample fraction is small but consistent, leaving the vast majority of trips unobserved.

At a more philosophical level, the question arises as to what constitutes the population being sampled. While it may be possible to interview a high proportion of drivers at some interview sites (if hardly practicable on major roads), such surveys rarely last more than one day. Models seek to represent long-term conditions, and no one day can represent annual conditions without some uncertainty. Ultimately, models seek to predict future conditions, which do not admit to complete observation by any current survey.

2.4.9 Negative binomial

Harris (1964) showed that, if instead of the cost coefficient λ having a single value throughout the population, the value is distributed according to a gamma function, then the resulting counts of trips will have a negative binomial distribution instead of the Poisson.

This hierarchical error structure is also used in accident modelling. In trip distribution modelling, variations within the population are usually treated by segmentation. In many cases segmentation incorporates useful information into the model, eg the effects of car ownership through household segmentation, and the different locations of employment, shopping and other land uses through segmentation by purpose. Brown (1982) showed that this can improve the fit of the overall model substantially.

2.4.10 Whittaker's formula

Whittaker (1979) derived an approximation for the accuracy of distribution model outputs from maximum likelihood theory. The formula gives an error factor for cell values estimated by fitting a distribution with a user-defined cost function to observed data.

$$\text{Error factor of estimate of } T_{ij} = \exp[\sqrt{(1/N_i + 1/N_j + 1/N_k - 2/N)}]$$

where

N_i and N_j are the observed trip end counts for the relevant origin and destination zones,

N_k is the total count of observations in the same cost band, k, as C_{ij}, and

N is the total count of observations.

This error factor works multiplicatively on a log scale for confidence intervals, so a 95% confidence interval would be

from $\quad \text{Estimate} \div \exp[1.96\sqrt{(1/N_i + 1/N_j + 1/N_k - 2/N)}]$

to $\quad \text{Estimate} \times \exp[1.96\sqrt{(1/N_i + 1/N_j + 1/N_k - 2/N)}]$

It is independent of factoring, so it can be applied on any scale – daily, hourly or sampled counts. The formula is easy to calculate; the distribution model does not have to be fitted. The error is due to the sampling of observations.

Gunn and Whittaker (1981) showed that this error factor, normally distributed, corresponds with the sampling errors in a simulation.

2.5 Combined modelling

Although trip distribution is a distinct stage in conventional transport modelling, it is rarely treated in isolation. Both Wilson's (1969) and Cochrane's (1975) key papers on distribution also considered mode split and Wilson included route split (assignment).

This pattern has continued with much interest in the interaction between different stages of modelling and the achievement of consistency between them. This has been fostered in the USA by requirements of its Intermodal Surface Transportation Efficiency Act of 1991, and studied by Boyce. It involves iteration through separate stages of the model, or modelling several stages jointly.

This study concentrates on distribution, but the following interactions with other stages of modelling are important considerations in trip distribution, or introduce topics for later chapters of this study.

2.5.1 Mode split and distribution

Mode split is, with distribution, one of the middle stages of the four-stage model. It represents travellers' choices of mode, typically between private (car) and public (bus/train), but also 'slow' or 'active' modes (walk or pedal-cycle) in the WTSM.

Mode choice is usually based on the same economic theory of user choice with random utility as described by Cochrane (1975, section 2.1.2) for distribution. Thus the probability of choosing a particular mode is

$$P(\text{mode}) = \exp(-\lambda_m \, \text{Cost}_{\text{mode}}) \; / \; \sum_{\text{mode}}(\exp(-\lambda_m \, \text{Cost}_{\text{mode}}))$$

where λ_m is a cost coefficient, reflecting unquantified utility in the modes, equivalent to λ_d in the choice of destination in distribution.

The theory dictates that mode split and distribution should be modelled in a hierarchy determined by the relative size of the coefficients of cost, λ_m and λ_d. Senior and Williams (1977) showed that although the order of the hierarchy may not affect fitted parameters or measures of fit greatly, it can affect the models' detailed response to policies, particularly those of restraint. This hierarchy has been the major influence on the form of London and Wellington demand models.

A small spreadsheet was built to examine the ordering of mode split and distribution under choice theory. The wrong ordering of mode split and distribution can give irrational results, with trips for one segment increasing in response to reduced costs in another segment, but this occurs mainly in extreme circumstances.

There are known to be ranges of data under which logsum coefficients larger than unity are consistent with utility maximisation, and do not result in irrational elasticities (K Train, pers comm). These have been examined by Herriges and Kling (1995; 1996) who cite further studies. Examples of estimated models with logsum coefficients over unity are given by Train et al (1987) and Lee (1999).

An earlier model of Wellington predicted that improved rail services from the Kapiti Coast would result in more car trips in the corridor, as well as more trips by train.

2.5.1.1 Composite cost

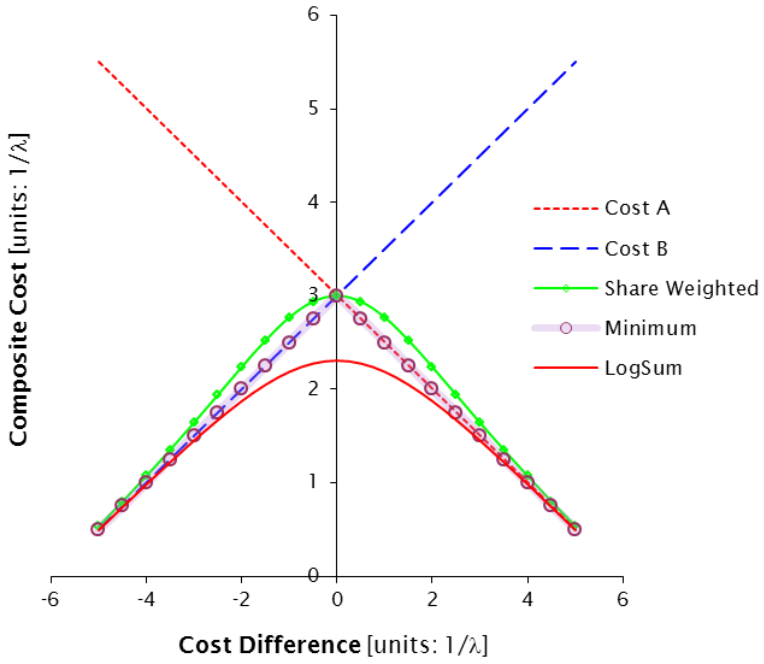
The issues revolve around the calculation of composite costs, also known as inclusive values. The first process in the sequence needs a composite of costs from choices modelled in later processes. If destination is to be chosen first, it needs to be based on the composite costs of all modes; if mode split is first, it is based on a composite cost across destinations. Economic theory indicates that the correct form of composite cost is the logsum

$$= -\text{Log}(\sum_{\text{choice}}(\exp(-\lambda\ \text{Cost}_{\text{choice}}))) \ / \ \lambda$$

This gives a composite cost lower than any of the individual costs, because the lowest cost is always an option, which is reduced further by the value of choice.

Figure 2.10 shows various forms of composite cost applied to two options, A and B, whose costs are shown as straight lines. The minimum simply follows the lower of the two lines. The logsum is less than the minimum. If the proportions choosing A and B are calculated from their costs using choice theory, these proportions can be applied to the costs to give the share weighted average. This is not the same as the logsum; in particular, it is greater than the minimum, contrary to the expectations of choice theory. Williams (1977) discussed other forms of composite cost and their desirable properties.

Figure 2.10 Composite costs



For N equal cost alternatives, the composite cost is

$$= -\text{Log}(N \times (\exp(-\lambda\ \text{Cost}))) \ / \ \lambda$$

$$= -\ (\text{Log}(N) + \text{Log}(\exp(-\lambda\ \text{Cost}))) \ / \ \lambda$$

$$= \text{Cost} - \text{Log}(N) \ / \ \lambda$$

Thus value of choice arising from unquantified elements increases with the number of choices and decreases with λ. It can result in negative composite costs.

The same effect is shown in figure 2.7 with the horizontal axis reversed because costs are negative utilities.

2.5.1.2 Order of modelling

The proper order of modelling is big choices, with greater unknown differences, before little ones: small λ before big λ. In mode choice this would be expected to give car vs public transport before bus vs train.

If $\lambda_d < \lambda_m$, then distribution should be modelled before mode split

If $\lambda_d > \lambda_m$, then mode split should be modelled before distribution, as in the WTSM except for HBW

If $\lambda_d = \lambda_m$, the order does not matter and mode split and distribution can be modelled jointly, as in the case of HBW in the WTSM

This is the order in which choices need to be modelled. It does not imply that travellers make decisions in the same sequence, or even that they use a sequential decision-making process. The order of modelling follows the amount of unquantified random value in travellers' choices. Theory does not determine the relative scale of these, and hence the modelling hierarchy; they are based on empirical observation alone.

2.5.1.3 Disaggregate modelling of discrete choices

Mode choice models are often calibrated differently from distribution models. Distribution models are calibrated on the aggregate number of observed trips for each zone-to-zone movement; each cell of the matrix is one record in the statistical analysis. Mode choice is calibrated on the chosen mode of each trip; each trip is one record in the analysis. With this disaggregate approach, individual circumstances can be set against each choice and their influence on the choice may be better distinguished.

The underlying theory of discrete choice is essentially that set out by Cochrane (1975) for destination choice and leads to the logit model. The simple binary logit model is a standard form of GLM, which can also be formulated to represent a multinomial logit model. However, these are only applicable where the level uncertainty is the same for all choices, ie $\lambda_d = \lambda_m$. Specialised programs for fitting to discrete choice data, such as Alogit and Biogeme, have been extended to models that allow different levels of uncertainty, $\lambda_d \neq \lambda_m$. The nested logit model is frequently used in transport modelling; the mixed logit model can represent nesting and many other forms.

These discrete choice methods can also be applied to distribution, as destination choice. However, there are only a handful of modes to choose from, but a large number of destinations – 225 zones in the case of the WTSM. This number of choices poses major computational problems for multinomial discrete choice models. McFadden (1978) showed that these can be overcome by sampling, but generally discrete choice models are applied to a limited set of destinations, eg:

- residential location
- shopping choice
- long distance travel
- tourism

and incorporate more detailed data on travellers and destinations than is generally available across an urban transport model for synthesising future scenarios. They have not been used widely for calibrating the trip distribution in major urban transportation models, but two notable applications in the UK have been reported during this study.

Daly and Ortuzar (1990) included sampled destination choice in an analysis of Santiago data, with the aim of improving the generalised cost function in the modal split.

2.5.1.4 PRISM

The policy responsive integrated strategy model (PRISM) provides a transport demand model for the West Midlands of the UK. RAND Europe (2004) fitted an advanced disaggregate choice model to household travel data, treating travel as tours rather than trips. Trip distribution and mode split were modelled jointly, using several household and person characteristics and incorporating choice of rail station and access mode, while trying to add time period choice. Fifty-nine parameters and constants were calibrated simultaneously in the commute model.

Not all the elements of conventional trip distribution are apparent in this complex model. The attractiveness of zones is scaled by their total employment for commuting, but there are no zonal balancing factors. Destination constants were added to improve the fit for Birmingham and Walsall in the commuting model. These may be acting as broader/coarser balancing factors, representing their accessibility effects. Mode-specific constants were also introduced to correct the modal split to some destinations and an intra-zonal constant allows for differences between modes in the treatment of intra-zonal movements; these may also act as balancing factors.

The consultants paid much attention to the ordering of distribution and mode split. They quoted likelihood statistics and their ratios, but their choice of model was often based on values of time and elasticities. Elasticities were calculated only for the 'right' order, so the existence of perverse elasticities with the 'wrong' order was not demonstrated. Some of the 'wrong' models have higher likelihood statistics.

Business and non-home-based trip distributions were calibrated from roadside interview data. Sophisticated corrections were applied to correct for sampling bias. These do not appear necessary given an understanding of partial matrix methods for conventional trip distribution, or equivalently the non-availability of destinations that do not involve crossing a screenline under discrete choice modelling. As a function of trip length, the correction term will be correlated with trip cost and its coefficient is liable to incorporate some of the real deterrence effects of cost. Corrections for multiple screenline crossings were also estimated in the calibration process rather than adjusted in the sampling rates.

2.5.1.5 Transport model for Scotland 2007

A much simpler joint distribution and mode split model was calibrated for TMfS07 from zonal matrix data (MVA Consultancy 2009). The Alogit program, usually applied to disaggregate data as in PRISM, was used to fit a nested logit model. This allows separate scales of uncertainty in distribution and mode split, fitting the ratio between them as Θ, the mode split spread parameter. Mode-specific Tanner cost deterrence functions and intrazonal K factors were fitted.

Trip end factors for the 712 zones were not fitted within the nested logit model, but were calculated outside and then included in the logit model as fixed values. Iteration between the nested logit model and trip end factoring is a sophistication of the Furness process. An audit report refers to the method as 'contraction mapping'.

No statistical measures of fit were given, but this might be because the observed trip matrices were of mixed provenance and could not be related to their original sample sizes.

2.5.2 Generation and distribution

Trip generation is the first stage of the conventional four-stage transport model. The number of trips produced by and attracted to each zone is estimated from zonal characteristics.

A doubly constrained trip distribution will replicate both the productions and attractions estimated in trip generation.

A singly constrained trip distribution replicates only productions; the input attractions are only a relative measure of attraction and are not necessarily matched by the output attraction trip end totals. These are a function of cost deterrents as well as input production trip ends and are scaled to match input productions overall.

There is thus an interaction between attraction functions and cost deterrence functions in calibration against observed data.

Daly (1982) developed a method for jointly calibrating trip attraction coefficients with destination choice. The disaggregate model had a non-linear component which needed a special programme to fit. The method applies only to attraction models with multiple explanatory variables, eg:

$$\text{Attractions} = a_1 \times \text{households} + a_2 \times \text{offices} + a_3 \times \text{shops} + \ldots$$

since it is the relative coefficients that may be affected. The scaling of the coefficient for a single explanatory variable is arbitrary in a singly constrained distribution, assuming no intercept term.

This form of interaction may be seen as an effect of accessibility on trip generation. Cochrane's (1975) partly constrained distribution (see section 2.1.3.4) provides a theoretical basis for such effects, but empirical evidence of such effects is often lacking.

2.5.3 Assignment and distribution

Assignment is the final stage of the conventional four-stage transport model. Paths are found between each origin and destination, the trips between those origins and destinations are loaded onto the network along those paths and the trips for all OD pairs are summed to give the total volume of traffic at each point in the network.

The choice of path can be based on the probabilistic random utility theory expounded by Cochrane (1975) for distribution (see section 2.1.3), and commonly used for mode split. This approach is most commonly used in public transport assignment, where it can provide sub-mode split (eg between bus and train), but it can also be applied to highway assignment, either analytically in Dial's methods or by randomisation in Burrell's.

However, in urban traffic the effects of congestion are often of greater concern, and these involve different methods such as incremental, volume-average or equilibrium assignment. All of these involve iteration, even with a fixed trip matrix. In congested assignment, costs of travel vary with the volumes of traffic on the network; in trip distribution the pattern of travel depends on the costs. There is thus interaction between trip distribution and congested assignment.

This interaction can be modelled by iteration through the whole of each model stage separately and successively. The WTSM takes this approach, including modal split in the iterative loop too. Evans (1976) proposed a joint model of distribution and assignment. The TransCAD software package can synthesise such a joint distribution-assignment, but only for a single purpose. It is understood that the ESTRAUS package, developed and used for Santiago, Chile, can handle the joint assignment of multiple user classes from different distributions.

No formal method of calibrating these distributions is offered beyond an empirical search for a value of the cost coefficient that reproduces observed network costs.

The result of these joint models is a distribution with an Exponential deterrence function, and an equilibrium assignment that meets Wardrop's criterion that all used paths have the same cost and all unused paths have a higher cost.

2.5.4 Aggregate – matrix and model estimation

Knowledge of the assignment process can be used to derive trip matrix information from vehicle counts on links or passenger counts on public transport routes. This aggregate data can be relatively cheap to collect, but matrices estimated from it are empirical observations of current movements around the network and do not readily provide forecasts of different demands under different circumstances.

The process is known as matrix estimation. In common with trip distribution, it seeks to generate trip matrices. In parallel with trip distribution, some methods appeal to the theory of maximum entropy, and advanced methods (MVESTM in the Trips/Cube suite – Logie and Hynd 1990) draw heavily on probability theory and maximum likelihood fitting.

MVESTM also provides for the calibration of a gravity model. There is a considerable history of estimating models from aggregate count data, and this is reviewed in chapter 8, which describes the calibration of trip distribution models from aggregate count data.

2.6 Calibration in practice

This review did not attempt to cover the practice of calibration comprehensively, but key aspects which have come to light are noted here.

2.6.1 Programs

Sen and Smith (1990, chapter 5) compared several algorithms for fitting distribution models.

2.6.1.1 GLIM

Among the methods they considered was GLIM, a statistical package for fitting GLMs, which shares its algorithms with Genstat, the program used in this research. Sen and Smith (1990) found that GLIM generally performed well, but was not as quick to compute as a modified scoring procedure. Both were notably 'even tempered', behaving predictably and well (p432).

GLIM was run on a Cray supercomputer, and as Sen and Smith said, it 'puts a heavy burden on the memory of even supercomputers' (p414), referring to work reported in 1992. One test with over a 100 zones could not be run in GLIM.

Sen and Smith (1990, section 5.9.2) found good small sample properties; however, the smallest scale of their sampling was 0.14 trips per cell, still generally an order of magnitude more than individual WTSM matrices (see table 2.2). They also found a good degree of robustness when the Poisson process was perturbed.

Sen and Smith (section 5.10.1) recommended the use of maximum likelihood procedures for fitting distribution models. Parameter estimates exist and are unique under very mild conditions, but lack diagnostics.

2.6.1.2 Independent programs

There are specialised programs for calibrating trip distributions. They are based on maximum likelihood, like GLMs, so their results should be the same and the same issues will arise in their statistical fit.

- LOGEST was developed by MVA for London Transportation Studies and also used in Scotland (CSTM3)
- John Bates' MAXL was used in Wellington
- George Skrobanski's GSLogitcal was used with OmniTrans in Dublin.

These all appear capable of simultaneously fitting complex models with both K and L factors and segmentation by household car ownership. Weighting by expansion factors does not appear to have been applied in the instances given above, although the issue is acknowledged.

Some transport modelling suites include calibration programs that are generally more limited.

2.6.1.3 Trips, in Cube

In the Trips suite, now incorporated into Citilabs' Cube, trip distribution is calibrated by the MVGRAM program in `Mode=1`. It can fit a Tanner (`GAMMA=T`) or Exponential deterrence function to a single full or partial (`PARMAT=T`) matrix. It maximises an objective function which is the form of a Poisson likelihood. For an ordinary Exponential deterrence function, the Power component or coefficient of log(cost) `X1` must be set to zero, and not allowed to default to unity.

MVGRAM can synthesise up to nine matrices simultaneously (eg by household car ownership, purpose or mode), with K factors, interpolating a user-defined (or empirical) deterrence function. However, these features cannot be calibrated.

2.6.1.4 VISUM

In VISUM, the KALIBRI function calibrates on a trip cost distribution, rather than a trip matrix. Trip ends and cost matrix are also input, and temporary trip matrices are synthesised giving empirical utilities by cost band. These are smoothed to Tanner or Exponential deterrence functions by regression, and the process is iterated. There are options of single or double constraint and weighting the cost bands by the number of observed trips.

2.6.1.5 TransCAD

The planning component of TransCAD includes calibration of gravity models. A set of empirical friction factors by cost band can be found, and a Tanner deterrence function is fitted to them by regression. Weighting the friction factors by trips gives a UTPS-like calibration.

Exponential and Power deterrence functions are fitted by iterative adjustment of their single parameters to replicate total travel cost. For the Power function, this will give a different result from maximum Poisson likelihood methods, which will replicate the total of the logarithm of costs (section 3.3).

There is a provision for generating K factors from a calibration, but these appear to be calculated after the fitting of the deterrence function, rather than as a simultaneous best fit. They may be zone-by-zone factors, simply the ratio between fitted and observed trip matrices. Selection appears to allow partial matrix methods. Tri-proportional models allow another set of constraints to be imposed.

2.6.1.6 EMME/2

EMME/2 does not offer a specific function for calibrating a trip distribution, but the three-dimensional option of the matrix balancing module can be used to fit an empirical deterrence function to cost bands. The cost bands are introduced as the third dimension matrix, together with the trip totals by band; the input matrix is flat, simply proportioned to the trip end totals. The empirical deterrence function is taken from the third dimension balancing coefficients, which have to be saved specifically.

Spiess' macro caligrav.mac calibrates an Exponential deterrence function to reproduce an average cost by successive approximation.

2.6.2 Models and their fitted parameters

Bly et al (2001) reviewed 24 models, including three in New Zealand, and list the trip distribution parameters for 12 of them in their table 8.2.

MVA (2005) list parameters from seven of the UK models they built.

The UK DfT (2006) cites both of these on its website for guidance on the conduct of transport studies (www.dft.gov.uk/webtag/). It favours the MVA list as having a provenance of 'rigorous estimation processes', particularly the ordering of mode split and distribution. Illustrative parameter values for destination choice are given in section 1.11.11 of the advice on 'variable demand modelling – key processes'.

The three New Zealand models reviewed by Bly et al are not the most recent major models. These are summarised in table 2.4. Cost coefficients λ for commuting (HBW) are taken from complex distribution/mode split models. In all models, the coefficients tend to be lower for longer trips. The Christchurch Transport Model ascribed much of the mismatch between household travel surveys and screenline counts to under-reporting rates, using the MVESTM matrix estimation package.

Table 2.4 Recent major New Zealand transport models

	Christchurch	Wellington	Auckland
Clients	NZTA, Environment Canterbury, Christchurch City Council	Greater Wellington Regional Council	Auckland Regional Council
Consultants	Traffic Design Group, MVA Asia	Sinclair Knight Merz, Becas	Sinclair Knight Merz, Becas
Advisors	John Bates		John Bates, Luis Willumsen
Principal analysts	Julie Ballantyne	David Ashley	Rohin Wood
Observation (census) year	2006	2001	2006
Data: numbers of interviews	2434 HIS RSI 8 external +13 internal sites 6092 on-bus	2538 HIS RSI 2 external (1 internal not used) 5079 train station	5221 HIS RSI 16 external sites 3444 bus 2260 ferry 4635 train
Zones	389 internal	225 internal + 3 external	517 (of which 1 external)
Externals	Separate distribution	Included in model	Included in model
Calibration software: distribution	Biogeme	MAXL (by John Bates in Delphi)	DMSion (by Rohin Wood in MatLab)
mode split	Biogeme	LimDep	LimDep
Synthesis by	Cube Voyager XCHOICE	EMME/2 Balmprod3.mac	EMME/3 ?
HBW distribution sensitivity: private λ (gen min)$^{-1}$	0.00101 to 0.117	0.0391 to 0.1175	0.06412 or 0.10664
PT λ (gen min)$^{-1}$	0.001 to 0.265	0.018 to 0.303	0.03457
HBW mode split sensitivity	$0.4 \times \lambda$	λ	λ
Household segmentation	Number of cars	Cars v adults	Tried, not needed
Geographic segmentation	3x3 sectors with short-medium-long trip lengths	6 levels of hierarchy	Broken stick at 9km

2.6.3 Scopes of K and L factors

The pattern of K and L factors is often based on administrative units, or natural boundaries such as rivers and major watersheds. They may also be determined by cordons and screenlines where there is data pointing to the need to improve the fit of a simple gravity model.

The hierarchical geographic segmentation of the Wellington model is described in section 4.4. It is used for both K factors, 'constants', and L factors, 'parameters', in different combinations for different household types and modes.

In the 1991 London Transportation Studies Model (MVA 1998), K factors were fitted as 'area specific constants' for segments defined by central, inner, outer and external sectors, and as 'river crossing constants' for trips crossing the river Thames, by direction. L factors were fitted as separate cost coefficients for movements:

- inbound to central London
- other inbound
- outbound
- orbital
- to external attractors.

Distribution and mode split were modelled on 529 districts, which were aggregations of the 1272 zones used for assignment.

Scottish models have used a sector system based on:

- Edinburgh
- Glasgow
- other.

Both the final Central Scotland Transport Model (CSTM3) and the 2005 Transport Model for Scotland (TMfS05) have separate K factors for each of the nine segments defined by these three sectors. CSTM3 also has separate L factors ('scaling parameters') for each of the nine segments, but TMfS05 has four separate 'sensitivity parameters', one for each of the three intrasector movements, and one for all intersector movements. In the more recent TMfS07 (see section 2.5.1.5) the only K factor is for intrazonal movements, as the demand model is applied incrementally by a combination of additive and multiplicative adjustment factors.

In the update of the Dublin Transportation Office Model following the 2002 Census, K factors were defined by socio-economic linkages following an examination of anomalous flows. Residential zones with a high proportion of residents in the AB groups were associated with office zones and the airport; zones with a low proportion were associated with mixed employment and manufacturing. Each such set of movements formed a separate segment with its own K factor. Movements within or between hinterland towns, or from external areas, also had their own K factors. A single 'β scaling parameter' was applied because there was 'not supporting data' for fitting separate L factors.

In the large complex models given as examples above, there is also segmentation by household car ownership. Different modes can also appear within a single distribution model, as in the WTSM, or be distributed separately, as for different purposes. The scopes of K and L factors are often amalgamated for the less common modes and purposes, where there is less information within which to draw distinctions. K factors are usually related to a reference level. A K factor cannot apply simply to a production or attraction zone, or a set of them, or it would be absorbed into the zone's balancing factor. L factors can be applied by zone (Emmerson 2008).

3 Analytical approach

This chapter draws together aspects of the analytical approach that underpin the methods of calibrating trip distribution models, or aid their interpretation. Some of these are simple or already known in some quarters; others apply to specific cases whose generality may be restricted. However, they do offer some understanding of the subject which is not readily available from transport modelling literature.

3.1 Calibration from a minimal 'four-square' set of data

A trip distribution model with an Exponential deterrence function can be written:

$$T_{ij} = A_i\, P_j \exp(-\lambda C_{ij})$$

A_i and P_j combine trip end totals and balancing factors. Other deterrence functions such as Power can be considered by substituting transforms like $C_{ij} = f(C'_{ij})$.

To estimate the cost coefficient λ, there is just sufficient information in a 2×2 matrix of trips T_{ij}, with corresponding costs C_{ij}.

Table 3.1 Minimal trip matrix

Prod\attr	1	2	Trip ends
1	T_{11}	T_{12}	T_{1*}
2	T_{21}	T_{22}	T_{2*}
Trip ends	T_{*1}	T_{*2}	T_{**}

Taking logs of the equation for each cell of the matrix gives:

$$\log A_1 + \log P_1 - \lambda C_{11} = \log T_{11}$$
$$\log A_1 + \log P_2 - \lambda C_{12} = \log T_{12}$$
$$\log A_2 + \log P_1 - \lambda C_{21} = \log T_{21}$$
$$\log A_2 + \log P_2 - \lambda C_{22} = \log T_{22}$$

Eliminating terms in A_i and P_j gives:

$$\lambda \times (C_{11} - C_{12} - C_{21} + C_{22}) = -\log T_{11} + \log T_{12} + \log T_{21} - \log T_{22}$$
$$= \log(T_{12} \times T_{21} / T_{11} \times T_{22})$$

This shows why there is no information about the coefficient λ if costs are linear,

ie when $(C_{11} - C_{12} - C_{21} + C_{22}) = 0$

or $C_{11} + C_{22} = C_{12} + C_{21}$

or $C_{ij} = C_{i*} + C_{*j}$

where C_{i*} and C_{*j} are the same for all * as with the shadow costs of the transportation problem in linear programming.

This will not usually be the case if the production and attraction zones are the same, as C_{11} and C_{22} will be intrazonal costs, and thus smaller than the interzonal costs, C_{12} and C_{21}. The attraction zones can differ from the production zones: the data could be any 'four-square' subset of the intersection of two rows and two columns of a larger matrix. Thus T_{11} and T_{22} need not be intrazonal trips, which are not observed by RSIs, and cost coefficients can be calculated from those parts of a trip matrix that are observed by such surveys.

The equation for the cost coefficient also shows the influence of the cross product term of the trips:

$$T_{12} \times T_{21} \; / \; T_{11} \times T_{22}$$

In classic testing of a contingency table this ratio is unity and its logarithm is zero, under the null hypothesis of no interaction. The cost coefficient can thus be seen to relate an interaction in trips to a differential in costs.

3.1.1 Sampling error

Assuming a reasonably well-specified model and known costs, the main source of error in λ arises from the sampling of observed trips T_{ij}.

$$\mathrm{Var}(\lambda) \times (C_{11} - C_{12} - C_{21} + C_{22})^2 \; = \mathrm{Var}(-\log T_{11} + \log T_{12} + \log T_{21} - \log T_{22})$$

Assuming that the sampling of trips T is a simple Poisson process

$$\mathrm{Var}(T) \; = T$$

so

$$\mathrm{Var}(\log T) \; = 1/T \quad \{ \text{ by linear approximation } \mathrm{Var}(f(x)) = (\mathrm{d}f(x)/\mathrm{d}x)^2 \mathrm{Var}(x) \; \}$$

For a matrix observed by, say, home interview, the trips in each cell of the matrix body, T_{11}, T_{12}, T_{21} and T_{22}, should be independent. Thus

$$\begin{aligned} \mathrm{Var}(-\log T_{11} + \log T_{12} + \log T_{21} - \log T_{22}) \; &= \mathrm{Var}(\log T_{11}) + \mathrm{Var}(\log T_{12}) + \mathrm{Var}(\log T_{21}) + \mathrm{Var}(\log T_{22}) \\ &= 1/T_{11} + 1/T_{12} + 1/T_{21} + 1/T_{22} \\ \mathrm{Var}(\lambda) \; &= (1/T_{11} + 1/T_{12} + 1/T_{21} + 1/T_{22}) \; / \; (C_{11} - C_{12} - C_{21} + C_{22})^2 \end{aligned}$$

Thus the accuracy of λ tends to depend on the size of the smallest count in the trip matrix and the differential in costs.

These simple calculations provide a useful view of the calibration process, demonstrating how a cost coefficient can be calculated from any four-square subset of a trip and cost matrix. Trip distribution models generally cannot fit exactly to larger matrices, since there is redundancy in the data. Much of the resulting difference between the model and observations can be ascribed to sampling error with a Poisson distribution. By maximising the likelihood under this distribution, trip distribution models of several forms, including those whose theoretical derivation is set out in the 'Literature review' (chapter 2), can be calibrated from large practical datasets using the statistical methods for fitting GLMs.

3.2 Generalised linear models

3.2.1 Regression

GLMs are founded in simple regression, finding the best line to go through a series of points.

$$\hat{Y} = mX + c$$

This can be extended into multiple dimensions, fitting more than one explanatory variable X, in a linear function of them.

$$\hat{Y} = a_0 + a_1 X_1 + a_2 X_2 + \ldots \; a_n X_n$$

The best values for the coefficients $a_0 \ldots a_n$ are found for the least squares of the error ε between the observed value Y and the linear predictor.

$$\varepsilon = Y - (\; a_0 + a_1 X_1 + a_2 X_2 + \ldots \; a_n X_n \;)$$

This gives the maximum likelihood provided ε is:

- normally distributed

- with constant variance
- without correlation between observations.

Computation is relatively simple and the results behave exactly according to known statistical distributions.

3.2.2 Generalisation

GLMs (McCullagh and Nelder 1989) extend linear modelling in two ways.

1 The error distribution no longer has to be normal. In particular, it can be Poisson to represent random counting processes.

2 A link function g() can be introduced between the expected value and the linear predictor

$$g(\hat{Y}) = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

In particular, when the link function is logarithmic, the additive relationship between the explanatory variables X in the linear predictor becomes multiplicative.

$$\text{Log } (\hat{Y}) = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

$$\hat{Y} = \exp(a_0) \times \exp(X_1)^{a1} \times \exp(X_2)^{a2} \times \ldots \times \exp(X_n)^{an}$$

Alternatively, they can become elements of generalised cost in an Exponential cost function.

$$\hat{Y} = \exp(a_0 + \lambda_1C_1 + \lambda_2C_2 + \ldots + \lambda_nC_n)$$

Computation requires an iterative process and the results approximate to the known distributions that apply to simple regression.

The main advantage of the GLM for calibrating trip distributions is that both the random (Poisson distribution) and systematic (log link) parts can be set to correspond exactly with the basic form of the distribution model. If applied to a synthesised matrix, a GLM can 'reverse engineer' it and recover the parameters used to build it – a highly desirable property for a calibration method.

3.2.3 Mixed models

Mixed models extend linear modelling by considering more than one level of variation. For example, there may be a variation between sites ε_s as well as an error between individual observations ε_i.

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n + \varepsilon_s + \varepsilon_i$$

In a sample, the apparent variability between sites is made up from components of both ε_i and ε_s. If their variances are σ^2_i and σ^2_s, and there are n observations per site, then ε_i contributes σ^2_i/n to the variance of each site mean, and the total variance in the sample between sites is

$$\sigma^2_s + \sigma^2_l/n$$

The variance between sites, σ^2_s, is usually much smaller than that between individual observations, σ^2_i. However, if n, the number of observations per site is large enough, it becomes the predominant component of error. Ordinary linear regression is not aware of σ^2_s and can ascribe undue importance to differences between sites from smaller values of σ^2_l/n.

The extra level of variation between sites can also be seen as a reduced variation within sites and thus a correlation between observations at the same site. Mixed models can fit a broad range of correlation patterns, including those between:

- time periods – time series analysis
- locations – geospatial analysis

- regression coefficients – repeated observations
- surveys or experiments – meta analysis.

Mixed models are also known as restricted (or residual) error maximum likelihood (REML) models. They relax the conditions of single, uncorrelated error terms in ordinary linear models.

They can also be thought of as analysis of variance (ANOVA) models, having multiple error strata, but without the need for balanced datasets of conventional ANOVA. Distinctions between strata are less clear in mixed models because of this lack of balance.

Galwey (2006) gives a clear introduction to mixed modelling with sample program code for the Genstat and R+ software packages.

3.2.4 Hierarchical models

In ordinary mixed models, all the error components are normally distributed. Hierarchical models generalise this, allowing a wider variety of distributions for the error terms, and allow a link function between the linear predictor and the expected value.

Hierarchical models can be fitted by a set of GLMs, but their fit has to be assessed on a particular scale. The h-likelihood statistics to do so are discussed by Lee et al (2006).

3.3 Maximum likelihood properties of Poisson log-linear models

In calibrating a trip distribution to observed data, a Poisson distribution of the observations can be expected from sampling error. This is a consequence of the observation process and is not dependent on any theory of trip distribution or destination choice. However, it does lead to some convenient and desirable properties in models fitted by maximum likelihood.

Under a Poisson distribution, the likelihood of observing T trips given an expectation of t trips from a fitted model is

$$t^T e^{-t}/T!$$

GLMs are fitted to maximise the product of such likelihoods over all observations, or equivalently minimise the deviance, D = −2log(likelihood) + K, scaled by K so the deviance is zero when the model fits the data, t=T. For a normal distribution, the deviance is the sum of squared residuals and is χ^2 distributed. When based on other distributions such as the Poisson, the χ^2 distribution of the deviance is approximate.

$$\begin{aligned} D &= -2\log\textstyle\prod(t^T e^{-t}/T!) + K \\ &= -2\textstyle\sum(T\log t - t - \log T!) + K \\ &= 0 \qquad \text{where } t=T \end{aligned}$$

Thus $$K = 2\textstyle\sum(T\log T - T - \log T!)$$

which is a constant function of the data T, unaffected by the fit of the model t.

$$\begin{aligned} D &= -2\textstyle\sum(T\log t - t - \log T!) + 2\sum(T\log T - T - \log T!) \\ &= 2\textstyle\sum T\log(T/t) - (T-t) \end{aligned}$$

Considering the minimisation of the deviance

$$dD/dt = 2\textstyle\sum(-T/t+1)$$

with respect to any parameter θ in the model t

$$dD/d\theta = \sum_{obs}[dD/dt \,.\, dt/d\theta]$$

summed over observations, usually trip matrix cells, but possibly aggregates such as segments or zonal trip ends

$$= 2\sum_{obs} [(1-T/t)\, dt/d\theta]$$

At the turning point $dD/d\theta = 0$, the average ratio of observations to modelled values T/t is unity when weighted by $dt/d\theta$. Note that this does not automatically reduce to equality in weighted sums, $\sum T(dt/d\theta) \neq \sum t(dt/d\theta)$, because t varies by observation and cannot multiply the summation on both sides. It is the case for a null model where t = constant.

For a trip distribution model of the form

$$t_{ij} = P_i A_j p_i a_j \exp(-\lambda Cost_{ij})$$

considering the minimisation of D with respect to a production trip end balancing factor, $\theta = p_i$

$$dt/dp_i = P_i A_j a_j \exp(-\lambda Cost_{ij})$$

$$= t_{ij} \,/\, p_i$$

$$dD/dp_i = 2\sum((t_{ij}-T_{ij})/p_i) = 0 \text{ at the turning point}$$

The factor p_i takes a constant value for production zone i and zero otherwise. Thus the maximum likelihood condition is that the sum of model values equals the sum of observations for production zone i, ie the model reproduces observed production trip ends for zone i. Similarly, the condition for a_j is that the model replicates observed attraction trip ends in zone j. Because all trip ends are replicated, total trips for the whole matrix are replicated.

Considering the minimisation of D with respect to the cost coefficient, $\theta = \lambda$

$$dt/d\lambda = P_i A_j\, p_i a_j\, Cost_{ij} \exp(-\lambda Cost_{ij})$$

$$= t_{ij} \times -Cost_{ij}$$

$$dD/d\lambda = 2\sum(T_{ij}-t_{ij})Cost_{ij}$$

$$= 2\sum(T_{ij}\times Cost_{ij} - t_{ij}\times Cost_{ij}) = 0 \text{ at the turning point}$$

Therefore the maximum likelihood condition for the cost coefficient λ is that the model replicates the observed total trip costs. If separate coefficients are fitted for linear components of cost, eg

$$\lambda_d Distance + \lambda_t Time$$

then the model replicates the total travel in terms of each component individually.

Similarly, it can be shown that:

- the Tanner deterrence function replicates the observed trip totals of both cost and log(cost)
- K factors, local constants, replicate total trips over their scopes
- L factors, local cost coefficients, replicate total trip costs over their scopes
- cost bands replicate total trips within each band
- under variable weighting, the weighted totals are replicated
- in partial matrices, total trips and trip costs are replicated over those observations that are available for the calibration.

3.3.1 Synthesis by GLM

Once the deterrence function f(cost) of trip distribution is calibrated, the synthesis of a trip distribution is relatively simple. An initial matrix is formed from the deterrence function $f(C_{ij})$ of the cost matrix C_{ij}; its rows and columns are then iteratively factored to the desired trip ends P_i and A_j by the Furness (or Fratar) process. When the matrix trip ends converge on P_i and A_j, the matrix will have the form of a trip distribution

$$T_{ij} = P_i p_i A_j a_j f(C_{ij})$$

where p_i and a_j are balancing factors which are the product of the row or column factors from all the iterations.

Trip distributions can also be synthesised by GLM, using a model of the form

$$T_{ij} = p_i a_j f(C_{ij})$$

where p_i and a_j combine trip ends and balancing factors. The value of the deterrence function $f(C_{ij})$ is known, so no coefficient is sought for cost or a function of cost, unlike calibration. Instead $f(C_{ij})$ is entered as an offset, leaving values to be found for the row and column factors a_i and b_j.

The dependent variable T_{ij} is the cell of any trip matrix that has the desired trip end totals P_i and A_j. From the maximum likelihood properties of the log-linear model with a Poisson error distribution developed above, fitting sets of dummy variables or factors for p_i and a_j causes the model to replicate the row and column totals of T_{ij}.

Thus T_{ij} could be:

- the observed matrix from which trip ends P_i and A_j are taken
- the 'flat' matrix, without cost deterrence effects, simply proportioned from the trip ends, $T_{ij} = P_i A_j / \Sigma T_{ij}$ (where $\Sigma T_{ij} = \Sigma P_i = \Sigma A_j$)
- any synthetic trip distribution fitting to the trip ends P_i and A_j , including a minimum cost solution to the transportation problem of linear programming.

Synthesising trip distributions by GLM is not particularly efficient; it can be slower than Furness and converge poorly.

3.3.2 Re-fitting synthetic trip distributions

Calibration to observed data will generally give different results if error distributions other than the Poisson are specified in the GLM, because the residuals between observed and fitted values will be weighted differently. However, when re-fitting a model to a synthetic trip distribution, there need be no such residuals, provided the model specification matches that used in the original synthesis. In these circumstances, a good fit can be achieved and the original coefficients can be recovered by GLMs using a variety of distributions. This has been demonstrated empirically for a few cases.

3.4 Elaboration of deviance under sparsity

The maximum likelihood methodology of GLMs allows the fit of models and the significance of their components to be assessed statistically. The measures and tests are those available for regression or analysis of variance, but involve some approximation where the error distribution is not normal. In particular, some approximations are poor when data is sparse.

Observed matrices used to calibrate trip distributions are sparse in that there are many more cells in the matrix than there are trips sampled in surveys. This means that most cells have no trips in them.

For HBW trips by car, the WTSM sampled 3045 trips in a 225 zone internal matrix. Excluding empty zones with no observed trip ends leaves 162 production zones x 194 attraction zones = 31,428 cells. This is 0.097 observations per cell, so at least 90% of cells must be empty. Other purposes and modes are likely to be sparser still.

Sparse log-linear models also occur in accident analysis and have been investigated by Maycock and Hall (1984) and Maher and Summersgill (1996).

The effects of sparsity on the deviance have been examined empirically and analytically by generating the expected deviance as the summation over the possible outcomes i counted in a single observation from a Poisson distribution with mean μ.

$$\text{Expected deviance} = \Sigma_i [\ (\text{Poisson Probability}(i|\mu)) \times \text{Deviance}(i|\mu)\ ]$$

The next two sections consider the expected deviance of a well-fitting model, ie the residual deviance. Section 3.6 considers the initial deviance in a model with an unrecognised systematic component, and the consequent change in deviance when that component is fitted as part of the model.

Figure 3.1 Poisson distribution

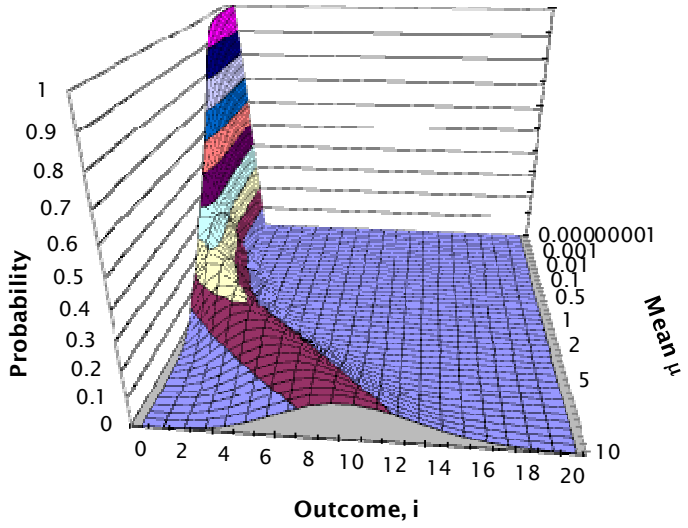


Figure 3.1 shows the Poisson probability distribution. The front of the graph shows a Poisson distribution with a mean of 10. The possible number of events i, from 0 to 20 occurrences, is shown across the graph. Their probability is given by the height of the plot. It reaches a maximum around 10 occurrences and its distribution is approximately normal since the mean μ is relatively large.

The mean μ diminishes along the axis going into the page. Its scaling is irregular, with a natural scale for large μ in the foreground expanding and then changing to a logarithmic scale for small μ in the background.

As μ diminishes, the bell-shape wave of the normal curve piles up against the limit of zero occurrences. As μ becomes very small, no occurrences become almost certain and any occurrences become highly improbable.

Figure 3.2 Deviance

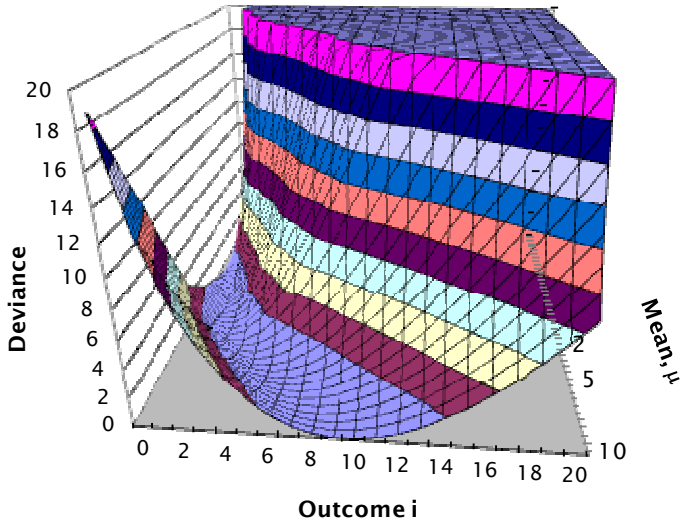


Figure 3.2 shows the deviance across the same horizontal dimensions. It is zero when the number of occurrences equals μ, and increases with the difference between them. It is the reversal of the Poisson probability, forming a valley where the probability forms a ridge. As μ becomes small, the bottom of the valley reaches zero occurrences, but the side of the valley continues to steepen as μ gets smaller.

Figure 3.3 Expected deviance – absolute

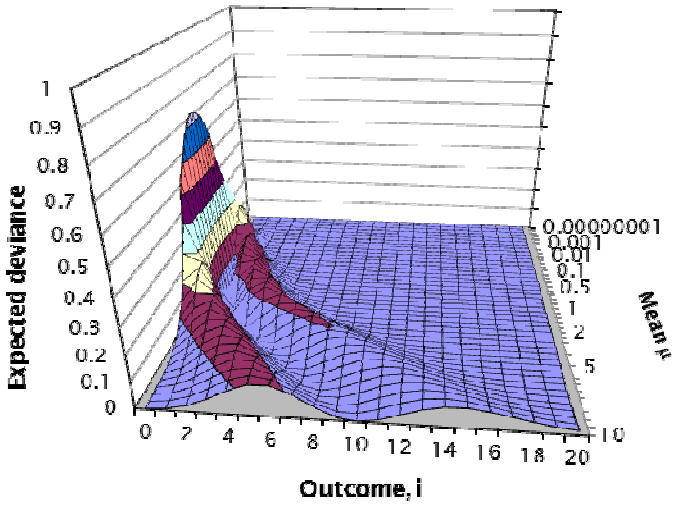
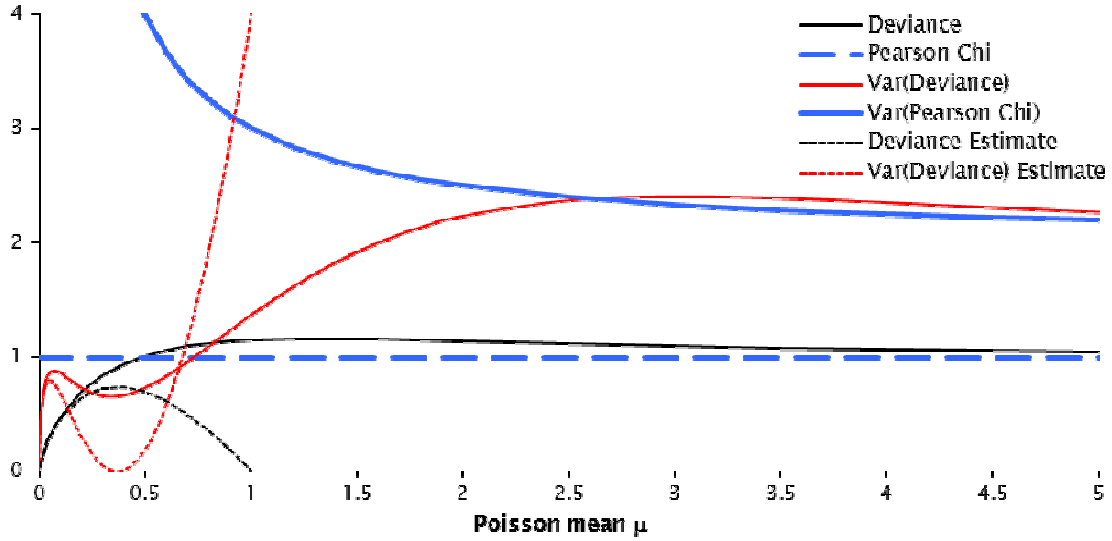


Figure 3.3 shows the expected deviance across the same horizontal dimensions. The expected deviance, shown vertically, is the product of the two previous plots and their contrary actions – probability decreasing away from the mean and deviance increasing. The result for larger μ in the foreground is two peaks, for fewer or more occurrences than the mean. They are not symmetrical, showing that μ=10 is still not large enough for the symmetrical normal distribution to be a good approximation for the Poisson.

As μ diminishes, the lower wave piles up against zero; having no occurrences when a small number is expected is the major source of deviance.

As μ becomes smaller still, no occurrences become more probable, and contribute less to the expected deviance. The upper wave, representing some occurrences when very few are expected, does not pile up to the same extent and the total expected deviance falls.

Figure 3.4 Measures of fit – natural scale



The expected deviance for any value of μ can be calculated by summing across the graph in figure 3.3. The variance can be calculated in the same way, and these are shown in figure 3.4 together with the mean and variance of the Pearson chi statistic.

This figure corresponds to Maycock and Hall's (1984) figure 10 and Wood's (2002) figures 5 and 6, and its derivation by Maher (1987, equation 7). A third turning point may be seen in the expected variance of the deviance.

Figure 3.5 Expected deviance – proportional

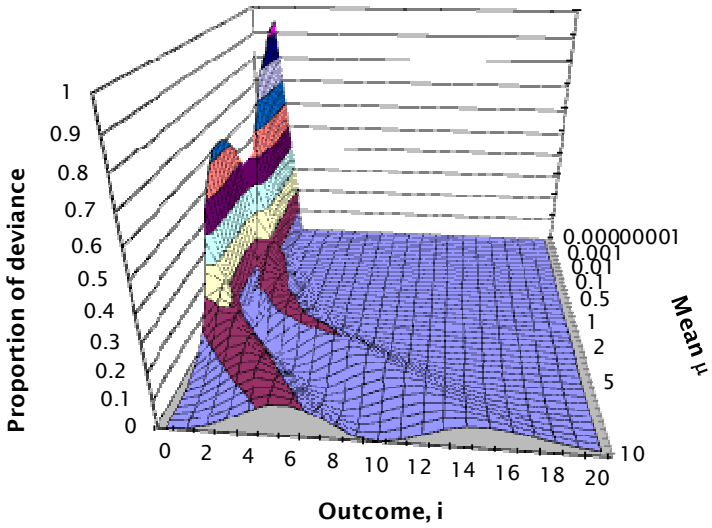


Figure 3.5 is the same as figure 3.3 except that the expected deviances are factored to sum to 1 across the graph, for a given mean μ. Thus it shows the relative contribution of each number of occurrences to the expected deviance.

The foreground is very similar to figure 3.3, since the deviance is expected to sum to 1 for large μ anyway.

As μ becomes small and the total deviance falls, this figure deviates from figure 3.3. As the proportion of expected deviance from zero occurrences falls, because the deviance is low for such a small μ, it is replaced by the expected deviance from one occurrence. The saddle point is around $\mu = 0.16$; 88% of the expected deviance is then derived equally from 0 and 1 occurrences. As μ becomes small, this proportion tends to $1-\mu$.

For low μ, most deviance arises from one occurrence. Multiple occurrences make little contribution because their probability is so low.

3.4.1 Analytical approximations

Given that most expected deviance is derived from just zero or one occurrence for low μ, only those two cases need to be considered to develop analytical approximations.

Table 3.2 Components of sparse deviance

Occurrences i	**Poisson probability, P**	**Deviance D**	**Expectation of deviance P × D**
	$e^{-\mu}\mu^i/i!$	**$2(iLog(i/\mu) - (i-\mu))$**	
0	$e^{-\mu}$	2μ	$2\mu e^{-\mu}$
1	$\mu e^{-\mu}$	$2(Log(1/\mu) - (1-\mu))$	$2\mu e^{-\mu}(Log(1/\mu) - (1 - \mu))$
Sum or for small μ, $e^{-\mu} \approx 1$, approx	$(1 + \mu)\, e^{-\mu}$ $(1 + \mu)(1 - \mu...)$ 1		$2\mu e^{-\mu}(Log(1/\mu) + \mu)$ $2\mu e^{-\mu}(-Log(\mu) + \mu)$ $-2\mu Log\mu$

Thus

$$\text{expected mean of the deviance} \approx -2\mu Log\mu \qquad \text{for small } \mu$$

Similarly, the expectation of the squared deviance is

$$e^{-\mu} \times (2\mu)^2 + \mu e^{-\mu} \times (2(Log(1/\mu) - (1-\mu)))^2$$

$$= 4\mu e^{-\mu} (\mu + (Log(1/\mu) - (1-\mu))^2)$$

$$= 4\mu e^{-\mu} (\mu + (- Log\mu - 1 + \mu)^2)$$

$$\approx 4\mu (Log\mu + 1)^2 \qquad \text{for small } \mu$$

The adjustment to find the sum of squares about the mean μ rather than about zero is $-\mu^2$. This is small in terms of the equation above, so can be ignored in estimating

$$\text{expected variance of the deviance} \approx 4\mu\, (Log\mu + 1)^2 \qquad \text{for small } \mu$$

These approximations are plotted on figures 3.4 and 3.6.

Figure 3.6 Measures of fit – logarithmic scale

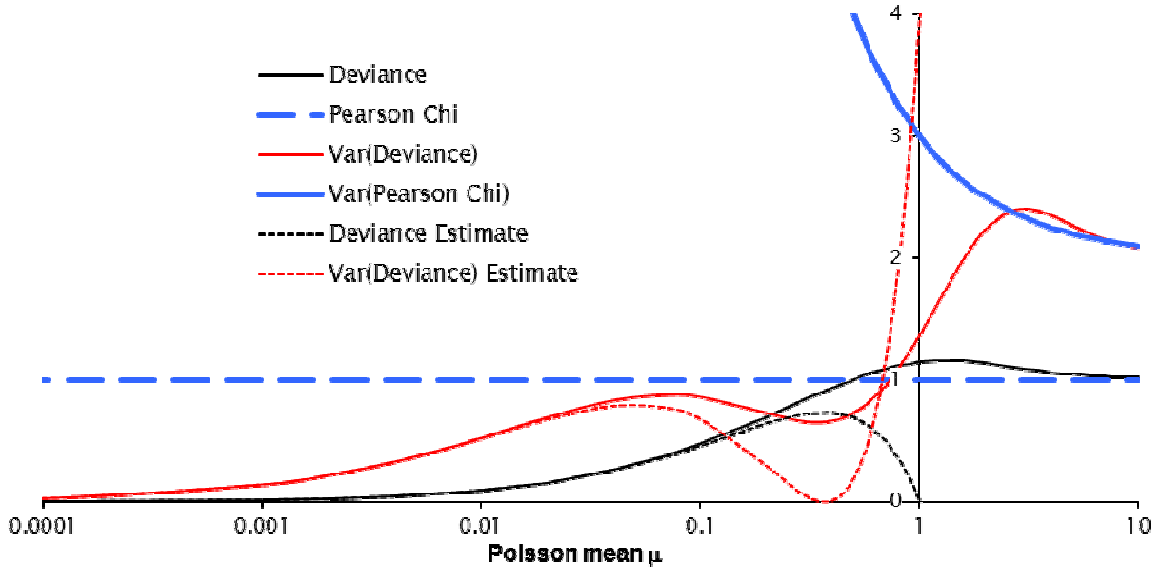


Figure 3.6 is a re-plotting of figure 3.4 onto a logarithmic horizontal scale to expand the small values of μ. The approximation for the mean looks good for $\mu<0.1$; that for the variance only for $\mu<0.01$. However, the variance approximation echoes two of the turning points in the exact value, so may be a useful component in some approximation with wider validity. Both approximations become poor as μ increases to 1.

3.5 Subdivision to a sparse dataset

Now consider a dataset with N events. They may be collisions in a crash model, or observed trips or home-workplace pairings in a trip distribution model.

The data is divided into R records. R may be relatively small if the units that the records represent are large – coarse zones or annual regional accident summaries; or large if the units are small – fine zones or monthly accident numbers for each junction. However, the total number of events N is the same.

Taking the simplest case of a homogenous rate of events across all units or records, the mean of the Poisson distribution for each record is then $\mu = N/R$.

3.5.1 Large mean μ

For large mean μ, N>>R, expectations of deviance are 1 for its mean and 2 for its variance (Wood 2002).

Table 3.3 Deviance characteristics – large mean

Deviance	Expected mean	Expected variance
Single record	1	2
Total of R records	R	2R
Mean of R records	1	$2R/R^2 = 2/R$

Adjustments for degrees of freedom in variances about estimated means are ignored.

Thus the relative standard error of the mean deviance is $\sqrt{(2/R)}$, and is dependent on the number of records.

By the central limit theorem, the distribution of the total deviance will tend to the normal as R becomes large; and so will the mean deviance, which is simply scaled from the total. The χ^2 distribution offers a closer approximation.

3.5.2 Small mean μ

For a small mean μ, N<<R, there will be many records with no events. The normalising influence of the central limit theorem will be more pervasive since R must be large for any useful value of N. Taking the expectations for deviance developed above:

Table 3.4 Deviance characteristics – small mean

Deviance	Expected mean	Expected variance
Single record	$-2\,\mu\text{Log}\,\mu$	$4\mu\,(\text{Log}\mu + 1)^2$
Total of R records	$-2R\mu\,\text{Log}\,\mu$	$4R\mu\,(\text{Log}\mu + 1)^2$
Mean of R records	$-2\mu\,\text{Log}\mu$	$4R\mu\,(\text{Log}\mu + 1)^2 / R^2$ $= 4\mu\,(\text{Log}\mu + 1)^2 / R$

Again adjustments for degrees of freedom are ignored.

The relative standard error of the mean deviance is then

$$\sqrt{(4\mu\,(\text{Log}\mu + 1)^2/R)} \;/\; -2\mu\,\text{Log}\mu$$

$$= -(\text{Log}\mu + 1) \;/\; -\sqrt{(R\mu)}\,\text{Log}\mu \qquad \text{taking the +ve square root: } \text{Log}\mu + 1 \text{ is -ve for small } \mu$$

$$= (1 + 1/\text{Log}\mu) \;/\; \sqrt{(N)} \qquad R\mu = N$$

$$= \sqrt{(1/N)}(1 + 1/\text{Log}\mu)$$

$$\approx \sqrt{(1/N)} \qquad \text{for very small } \mu$$

This shows that the main determinant of the mean deviance's variability is N, the number of events in the dataset. The amount of information represented by the N events is thus most important; how thinly they are spread amongst the records, much less so.

This expected relative standard error in the mean deviance is about the same as that in estimating the mean from a count of N events.

3.5.3 Conclusions

This sparse deviance is clearly no longer χ^2 distributed. It no longer has the same expectation of mean or variance, or the same relative standard error relating them. Therefore the mean residual deviance cannot be used as the denominator in an F test for a change of deviance.

The mean residual deviance is expected to be less than unity; this is not necessarily indicative of underdispersion. Therefore standard errors and t ratios should not be automatically scaled by the mean residual deviance (ie by setting `DISPERSION=*` in Genstat). The ability to detect over- or under-dispersion in a sparse mean residual appears to be determined by the number of observations in the dataset.

The mean residual deviance is more difficult to interpret under sparsity because its expectation depends on the fitted means μ, or their distribution, which will differ between models. For any large μ, the expectation is unity, so any model's residual deviance can be tested against unity. Under sparsity, there may be alternative models with different expected deviances – see sections 4.8.1 and 8.7.3.5.

3.6 Change in deviance

Although the properties of the residual deviance are affected by sparsity, models can be compared by differences in their deviances, particularly the change in deviance as a variable is added. This change in deviance is more robust, as is demonstrated for the following case.

The expectation of deviance can be written in terms of i occurrences from a single observation of a Poisson process with mean μ:

$$\Sigma_i[\ (\text{Poisson probability}(i|\mu)) \times \text{deviance}(i|\mu)\ ]$$

$$= \Sigma_i[\ P(i|\mu) \times 2(i\text{Log}(i/\mu) - (i-\mu))\ ]$$

$$= 2\Sigma_i[\ P(i|\mu) \times (i\text{Log}(i) - i\text{Log}(\mu) - (i-\mu))\ ]$$

$$= 2(\ \Sigma_i[\ P(i|\mu) \times i\text{Log}(i)] - \mu\text{Log}(\mu) - (\mu-\mu)\ ) \qquad \text{since } \Sigma_i[\ P(i|\mu) \times i] = \mu$$

$$= 2(\ \Sigma_i[\ P(i|\mu) \times i\text{Log}(i)] - \mu\text{Log}(\mu)\ )$$

However, for an imperfect model, the fitted value m is in error by a factor of x, so $m = x\mu$ is used to calculate the deviance, but the probabilities are still determined by the true mean μ.

$$\Sigma_i[\ P(i|\mu) \times 2(i\text{Log}(i/m) - (i-m))\ ]$$

$$= 2\Sigma_i[\ P(i|\mu) \times (i\text{Log}(i) - i\text{Log}(m) - (i-m))\ ]$$

$$= 2\Sigma_i[\ P(i|\mu) \times (i\text{Log}(i) - i\text{Log}(x\mu) - (i-x\mu))\ ]$$

$$= 2(\ \Sigma_i[\ P(i|\mu) \times i\text{Log}(i)] - \mu\text{Log}(x\mu) - (\mu-x\mu)\ ) \qquad \text{since } \Sigma_i[\ P(i|\mu) \times i] = \mu$$

$$= 2(\ \Sigma_i[\ P(i|\mu) \times i\text{Log}(i)] - \mu\text{Log}(\mu) - \mu\text{Log}(x) - \mu(1-x)\ )$$

From above, the expectation of deviance for a true model is:

$$2(\ \Sigma_i[\ P(i|\mu) \times i\text{Log}(i)] - \mu\text{Log}(\mu)\ )$$

and the difference is

$$2(\qquad\qquad\qquad\qquad - \mu\text{Log}(x) - \mu(1-x)\ )$$

$$= 2\mu\ ((x-1) - \text{Log}(x))$$

For x=1, this is 0, which is the deviance expected when modelled values equal observations.

The true mean μ can be divided among many records, not necessarily of the same size, but as long as the model is in error by the same factor x, the excess in the total deviance will be the same. For example, if the observation with mean μ is disaggregated into two records with means μ_1 and μ_2, where $\mu_1 + \mu_2 = \mu$, then the change of deviance for those two records is

$$2\mu_1\ ((x-1) - \text{Log}(x)) + 2\mu_2\ ((x-1) - \text{Log}(x))$$

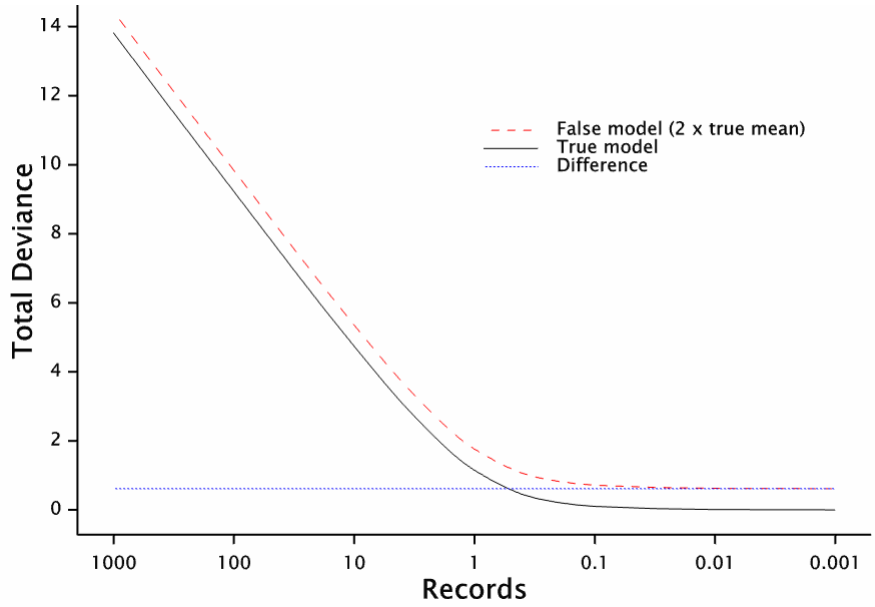
$$= 2(\mu_1+\mu_2)\ ((x-1) - \text{Log}(x))$$

$$= 2\mu\ ((x-1) - \text{Log}(x))$$

ie the same as the change of deviance for a single aggregate record.

The following plot is for $\mu = 1$, $x = 2$, spread equally across different numbers of records. This gives a difference in deviance of $2 \times 1 \times ((2-1)-\text{Log}(2)) = 0.6137$.

Figure 3.7 Deviance of true and false models



As the data becomes sparse, spread over many records towards the left of the plot, the total deviance increases for both the true and false models. However, the increase in deviance does not keep up with the number of records, so the mean deviances decrease with sparsity, as shown in figure 3.7 for a true model.

The difference in deviance remains a constant, suggesting the change in deviance offers a robust test against the χ^2 distribution irrespective of sparsity. However, the difference reduces as a proportion of the total or mean deviance of the false model with sparsity. Section 3.5.2 suggests that the expected relative standard error of the mean deviance is fixed by the number of observations, N, at $1/\sqrt{N}$. Thus as a dataset is divided more thinly, the power to detect an error in fit from the residual mean deviance diminishes.

The changing ratio between the two deviances and the difference between them implies that F or R^2 statistics will also vary with the number of records.

Although the cases considered here are simplistic and not universal, they are consistent with the advice on the interpretation of deviances in GLMs given by McCullagh and Nelder (1989, pp36 and 119) and Payne et al (2009, p304).

3.7 Loss of deviance change with aggregation

Later chapters analyse data disaggregated from production zones to households, persons or trips, or aggregated to screenline counts. This section considers the effect of aggregation on the change in deviance if the error in the model, x, is not consistent across the dataset.

If the ratio x of the estimated mean m to the true mean μ takes two values dividing the dataset into

$$m_1 = x_1\mu_1$$

and $$m_2 = x_2\mu_2$$

then if the deviances of the subsets are calculated separately, the change in deviance is

$$2\mu_1 ((x_1-1) - \text{Log}(x_1)) + 2\mu_2 ((x_2-1) - \text{Log}(x_2))$$

$$= 2\{ (\mu_1x_1+\mu_2x_2) - (\mu_1+\mu_2) - (\mu_1\text{Log}(x_1)+\mu_2\text{Log}(x_2)) \}$$

If the datasets are treated in aggregate, the overall ratio

$$x_* = (m_1+m_2) / (\mu_1+\mu_2)$$

$$= (\mu_1 x_1+\mu_2 x_2) / (\mu_1+\mu_2)$$

which is the average of the ratios weighted by the true means.

The change in deviance for the aggregated dataset is

$$2(\mu_1+\mu_2)\{(x_* - 1) - \text{Log}(x_*)\}$$

$$= 2(\mu_1+\mu_2)\{((\mu_1 x_1+\mu_2 x_2)/(\mu_1+\mu_2) - 1) - \text{Log}((\mu_1 x_1+\mu_2 x_2)/(\mu_1+\mu_2))) \}$$

$$= 2\{ (\mu_1 x_1+\mu_2 x_2) - (\mu_1+\mu_2) - (\mu_1+\mu_2)\text{Log}((\mu_1 x_1+\mu_2 x_2)/(\mu_1+\mu_2)) \}$$

The change in deviances calculated separately is greater than calculated in aggregate by

$$2\{ (\mu_1+\mu_2)\text{Log}((\mu_1 x_1+\mu_2 x_2)/(\mu_1+\mu_2)) - (\mu_1\text{Log}(x_1) + \mu_2\text{Log}(x_2)) \}$$

Dividing through by $2(\mu_1+\mu_2)$ gives

$$\text{Log}((\mu_1 x_1+\mu_2 x_2)/(\mu_1+\mu_2)) - (\mu_1/(\mu_1+\mu_2) \text{Log}(x_1) + \mu_2/(\mu_1+\mu_2) \text{Log}(x_2))$$

This is the difference between the log of the weighted average of ratios and the weighted average of the logs. Since the log function is concave ($d^2\text{Log}(x)/dx^2 = -x^{-2}$), this is always positive (for all x, μ, m +ve).

Thus aggregation always reduces the change in deviance (except for all x equal, as discussed in the preceding section)

This finding is based on expectations and is thus subject to perturbation by random effects. It considers the reduction in deviance from a mis-specified model to a true one. The true model may not be known, but a good model may be close to it, with little extra deviance, so the finding is very likely to hold true for the change of deviance from a poor model to a reasonably good one. This is the case for the change of deviance with the introduction of a cost coefficient; a flat model without trip distribution effects is a poor model compared with a trip distribution model.

3.8 Empty zones

With sparseness, some whole zones have no trip ends observed in them, either as productions or attractions. Their whole rows or columns in the trip matrix all have zero cells.

In a doubly constrained trip distribution, there are no trips to be distributed from or to these empty zones, and they can contribute no information to the calibration process. A model that includes empty zones can fit them with a value of zero for the trip end balancing factor, so all the fitted values for the row or column will be zero and will match the observations exactly.

These exact fits contribute nothing to the total deviance. If included in the number of observed cells, they increase the degrees of freedom and hence reduce the mean deviance. This measure of fit is improved by the inclusion of the perfectly fitted observations, although they contribute nothing to the information about trip distribution. More zones can appear to produce a better fit from the same survey data.

In a multiplicative model, the trip end balancing factors cannot be zero, but will be very small numbers, and the differences from zero will make a negligible contribution to the total deviance. Genstat does not appear to suffer any computational difficulties, but practical engineers may not want statistics packages to explore the mathematics of infinitely small numbers while they are calibrating a trip distribution. The small numbers may also generate warning messages.

To avoid such effects, empty zones and the cells in their corresponding rows and columns can be removed from the dataset before calibration.

Note that even after empty zones have been removed from the matrix, there will still be many cells with observations of zero; but there will be no complete rows or columns of them.

3.9 Zero cells

Apart from empty zones, matrix cells whose movements have been observed, but for which the observation is zero, are an important part of the dataset. These 'zero cells' contain useful information, that the volume of the movement is probably small.

They must be distinguished from null cells for unobservable movements, which cannot be observed. These typically arise from RSIs, where no reasonable route between two zones passes through the interview site or screenline. The movement between the two zones is then unobservable; the volume may still be large, but the survey offers no information about it.

The simple analysis of a four-square data subset in section 3.1 suggests that a cost coefficient cannot be calculated if any of the cells are zero. However, under the statistical approach of log-linear modelling, zero observations are recognised as a likely outcome of a small but finite trip probability.

Excluding zero cells from the dataset has serious consequences, as table 3.5 demonstrates.

Table 3.5 Treatment of zero cells when fitting trip distribution

	Observed	Fitted		
		treatment of matrix cells with observed count=0		
		excluded	included, equal weight	included, no weight
Trips where count > 0	183,216	183,216	59,716	183,216
Trips where count = 0	0	~	123,500	3,300,309
Total trips	183,216	183,216	183,217	3,483,525
Travel where count > 0	4,451,140	4,451,145	979,643	4,451,145
Travel where count = 0	0	~	3,471,518	185,905,626
Total travel (trips × cost)	4,451,140	4,451,145	4,451,161	190,356,771
Average trip cost (gen min)	24.29	24.29	24.29	54.64
Cost coefficient		0.0104	0.0638	0.0104

The first column of results for fitted models shows the effects of excluding zero cells. Following the findings of section 3.3, the model replicates the observed trips and travel (trips×cost) over the cells included in the analysis, ie only those where trips have been observed.

The second column shows the proper formulation, with zero cells included with equal weight. Total trips and travel are again replicated, but in this case spread across all the cells from which they were observed.

The third column shows the effect of including zero cells, but with zero weight. They are thus ignored in the calibration of the model, which estimates the same cost coefficient as when the zero cells are excluded. However, this model also fits values for the zero cells, grossly exaggerating the total trips and travel across all the cells. This can be seen as extrapolation from the non-zero cells to the zero cells, or as applying partial matrix methods to the non-zero cells alone.

In a sparse matrix, with low probabilities of trips in many cells, the counting of trips in a few of those cells is largely a matter of chance. Analysis of those cells alone introduces great bias.

Zero cells must be included not only in the dataset, but also in the weighting scheme.

3.10 Null cells – partial matrices

Zero cells must be distinguished from null cells representing unobservable movements. Null cells typically arise from roadside interview surveys, where no reasonable route between two zones passes through an interview site or screenline. The movement between the two zones is then unobservable; the volume may still be large, but the survey offers no information about it.

Since they contain no information about the movement, the data record can and should simply be omitted from the dataset for calibrating trip distribution by GLM. This is possible if there are still 'four-square' sets of observed cells (section 3.1) after the null cells have been eliminated. Observed trip matrices with null cells are known as partial matrices, and methods of calibrating trip distribution models from them are well established, if not widely understood.

These partial matrix methods can infill null cells with estimates from the distribution model. This may be achieved in GLMs by including the null records with zero weight (and a dummy dependent Y value). In Genstat, the dependent Y variable can be set to a missing value.

There are no null cells in the fully observed internal trip matrix from the WTSM HIS used in the main part of this study. When external↔internal trips from the RSIs on the study area boundary are included (section 4.12), external↔external movements are omitted from the analysis. In WTSM these movements were included with a cost of 999 generalised minutes.

3.11 Weighting

The statistical tools for assessing the fit of models and the accuracy of their coefficients depend on specifying the errors in the input data. In ordinary regression or analysis of variance based on the normal distribution, the scale of error can be estimated internally from the residual mean square errors. In log-linear models, the scaling of Poisson errors can be checked against the equivalent residual mean deviance, which should be unity for a well-fitting model with large means. However, section 3.5.2 shows that for sparse data this measure is difficult to interpret, and it may be insensitive to miss-specification (sections 4.8 and 8.7.3.5).

The prime source of error in the input data is from random sampling. Sampling error depends on the number of observations. When these are expanded to estimate the whole population, the sampling error is also increased; it is more than the error expected from a sample size of all the expanded observations. Weighting in accordance with the sampling rate can compensate for this increased error when working with expanded data.

3.11.1 Variable sampling rates

Most theoretical treatments of trip distribution do not consider sampling issues, or only the simple case of a constant sampling and expansion factor over the whole dataset. The RDMVAR programme in the ROADWAY suite provides a hierarchical structure for practical survey expansion which also allows the calculation of accuracies or weights.

In practice, sampling rates vary. The final control totals for expanding household surveys may only become available when census data is processed, well after the travel survey if it is concurrent with a new census; the survey sampling has to be planned from old census data or its projections. Refusal rates can vary, resulting in complex expansion schemes to maintain balance in household and person types. Roadside interviews and on-board public transport surveys typically provide higher rates than can be afforded in household interviews. Sampling rates will tend to vary with traffic volume since there is a limit to the number of interviews a survey team can complete in a roadside interview bay or in a passenger vehicle. All of these variations are found in the WTSM.

Two issues arise from varying sampling rates.

1 Calculation of weights for a set of observations within which sampling rates and hence expansion factors vary.

2 Identifying those scopes within which a single weighting factor applies and between which the weights vary.

Only variability from the sampling of observed trips is considered in forming weights and is presumed to take the Poisson distribution. Errors may also occur in the expansion factors, but these are generally based on larger numbers than the observed trips and hence involve smaller errors (CN7, end).

In a simple sampling scheme where the expansion factor is constant, the weight is its inverse, reducing the error model back to the count of observations. Where the sampling rate varies, the over-sampled part becomes more accurate and the under-sampled part less so. When combined, the overall accuracy is no longer the same as would be achieved by constant sampling at the overall rate. The overall error is increased by an inflation factor reflecting inefficiency in the sampling scheme. The effective sample size in the error model is less than the actual count of observations and the weight is no longer the inverse of the overall expansion factor.

3.11.1.1 Examples

A movement is split between two parts A and B, which are observed separately. These might be two crossings of a screenline formed by a river with a roadside interview point on each bridge; two public transport services in the same corridor with surveys on board each; or two types of household with different response rates.

Let the total volume be 2000 trips with 200 observations overall and consider different divisions between the two parts, A and B, starting with equal division of both volume and observations.

Table 3.6 Weighting for equal volumes, equal observations

	A	B	Overall
Volume (*Trips*)	1000	1000	= 2000
Observations (*Count*)	100	100	= 200
Expansion factor	10	10	10
Sampling variance of observations	100	100	~
Sampling variance of expanded observations (*Var*)	10,000	10,000	= 20,000
Variance from a sample size of whole volume	1000	1000	2000
Weight	0.1	0.1	0.1

'=' Overall values calculated as sum of parts A and B; otherwise calculated from other overall values, above

The values in the table are calculated as follows:

Expansion factor

= volume/observations

Sampling variance of observations

= observations, as a Poisson process, ignoring hypergeometric effects of finite populations

Expanded observations (estimate of volume)

= observations x expansion factor

so sampling variance of expanded observations

= sampling variance of observations x (expansion factor)^2

= volume^2/observations for parts with constant sampling.

Overall volume, observations and sampling variance of expanded observations can be summed across parts. Thus the overall sampling variance of expanded observations is Σ(volume^2/observations)

The weight is the factor by which the accuracy from the sample of observations is less than expected if the sample size were the whole volume.

The variance from a sample size of the whole volume

= Σvolume, again for a Poisson process.

Thus weight

= variance from a sample size of whole volume/sampling variance in expanded observations

= Σvolume / Σ(volume^2/observations)

= 1/(volume/observations) = 1/expansion factor for constant sampling.

If the expansion factor is constant across all the samples, the overall weight is its inverse. However, if the expansion factor varies, the weight is greater than the overall expansion factor (=Σvolume/Σobservations). This inflation represents the inefficiency of the sampling scheme and depends on the variation in expansion factors. This can be seen in the first two of the three cases below.

Table 3.7 Weighting for equal volumes, unequal observations

	A	B	Overall
Volume (*Trips*)	1000	1000	= 2000
Observations (*Count*)	150	50	= 200
Expansion factor	6.67	20	10
Sampling variance of observations	150	50	~
Sampling variance of expanded observations (*Var*)	6667	20,000	= 26,667
Variance from a sample size of whole volume	1000	1000	2000
Weight	0.15	0.05	0.075

'=' Overall values calculated as sum of parts A and B; otherwise calculated from other overall values.

This unequal sampling inflates the sampling error by 33%, with a reduced weight.

Table 3.8 Weighting for unequal volumes, equal observations

	A	B	Overall
Volume (*Trips*)	500	1500	= 2000
Observations (*Count*)	100	100	= 200
Expansion factor	5	15	10
Sampling variance of observations	100	100	~
Sampling variance of expanded observations (*Var*)	2500	22,500	= 25,000
Variance from a sample size of whole volume	500	1500	2000
Weight	0.2	0.067	0.08

'=' Overall values calculated as sum of parts A and B; otherwise calculated from other overall values.

This unequal sampling inflates the sampling error by 25%, with a reduced weight.

Table 3.9 Weighting for unequal volumes, unequal observations, but equal sampling rates

	A	B	Overall
Volume (*Trips*)	500	1500	= 2000
Observations (*Count*)	50	150	= 200
Expansion factor	10	10	10
Sampling variance of observations	50	150	~
Sampling variance of expanded observations (*Var*)	5000	15,000	= 20,000
Variance from a sample size of whole volume	500	1500	2000
Weight	0.1	0.1	0.1

'=' Overall values calculated as sum of parts A and B; otherwise calculated from other overall values.

With equal sampling rates there is no inflation despite differences in the volumes and sample sizes.

3.11.1.2 Data handling

In practice, expanded matrices are built by summing expansion factors over survey records into an accumulator named, say, *Trips*:

Trips = Σ(expansion factor), summed over individual trip records

= Σ(volume/observations)

= Σ(volume/1) = Σ(volume) since each record is one observation.

To calculate the weight, the squares of the expansion factors can be summed in parallel into another accumulator, *Var*:

Var = $\Sigma(\text{expansion factor})^2$, summed over individual records

= $\Sigma(\text{volume/observations})^2$

= $\Sigma(\text{volume}^2/\text{observations})$ since each record is one observation

= sampling variance of expanded volumes

Thus weight = *Var*/*Trips*

The number of observations can also be aggregated in parallel into an accumulator, *Count*:

Count = $\Sigma 1$, summed over individual records

giving an effective expansion factor of *Trips*/*Count*, which can be compared with the inverse of weight.

The calculation of weights can thus parallel the usual expansion process in matrix building.

3.11.2 Scope

3.11.2.1 Cells

A weight can be calculated for any matrix cell for which there are observations with expansion factors. However, there are typically many cells without observations. The weight applied to these zero cells affects the trip distribution model, since it reflects the importance of there being no observations – whether it could arise from a large actual volume due to a small sampling fraction, or whether, with a large sampling fraction, there is a strong implication of a small actual volume.

3.11.2.2 Wider scopes

Weights can be calculated over wider scopes than individual matrix cells and then applied to all cells, or all zero cells within that scope. The objective is to define scopes within which the probability of sampling a trip is equal so a common weight can be applied properly across each whole scope. It is desirable that the scopes are as large as possible, so that there is at least one observation in each and preferably a large number to reduce random effects when calculating weights from expansion factors of observations.

3.11.2.3 WTSM expansion scheme

Ideally, the weighting scheme is developed from the survey sampling and data expansion scheme. In the WTSM household survey this is complex, with separate schemes for adjusting by household and by person. The scopes vary, not always according to the documentation, but are generally based on census area units. Since zones may be split across area units and vice versa, it is very difficult to identify sets of matrix cells with consistent expansion factors beyond the zonal scopes described below.

3.11.2.4 Production zones

By definition, the production zone of home-based trips is also the location of the home. The sampling of households in a zone determines the probability of observing trips produced from the zone. The productions for each zone or matrix row can thus define a separate scope for weighting. This is the closest practical approximation to the scopes of the WTSM survey expansion scheme. If no trip productions are observed from a zone, it can be omitted as an empty zone, so all remaining zones will have observations with expansion factors from which a weight for the zone can be calculated.

3.11.2.5 All

The simplest scope is the whole matrix, with a single weight. The variance calculations in RDMVAR for observed zero cells (CN9.3) are equivalent to this scope.

3.11.2.6 Comparison

Table 3.10 compares the results of fitting a trip distribution model with different scopes of weighting. The weighting includes the penalty for varied sampling within each scope. In the first row, zero cells have no weight and the results differ greatly from those of all other schemes, which are broadly similar. The individual weighting for non-zero cells, with the single weight from all observations for zero cells, stands out slightly.

Table 3.10 Scope of weighting

Scope	Cost coefficient		Total trips	Average cost gen min	Correlation of trip ends	
	estimate	standard error			production	attraction
Cell (0)	0.01173	0.00134	3,146,683	49.63	0.2716	0.2874
Cell (zone)	0.06331	0.00141	186,019	24.87	0.9991	0.9909
Cell (all)	0.06070	0.00141	197,208	26.82	0.9241	0.9902
Production zone	0.06355	0.00142	183,219	24.64	1	0.9928
All	0.06377	0.00153	183,217	24.29	1	1

(scope for weighting zero cells is shown in brackets)

The model has an Exponential deterrence function and the fitted coefficients of cost are shown. In a well-specified model, proper weighting should not affect the estimate significantly, but should minimise the errors. From the standard errors, the single weighting from all observations is not so efficient; more complex weighting improves this accuracy by about 8%, equivalent to a useful increase in sample size of about 16%.

A consequence of varied weights in fitting a log-linear model is that it no longer replicates the simple totals of expanded trips, trip ends, or travel (trips×cost, with an Exponential deterrence function). Instead, it reproduces weighted totals. The simple total of trips and average of cost is shown in the table. Because there is no variation in weighting in the 'all' scope, it reproduces the observed values closely.

Correspondences between the observed and fitted trip end totals (unweighted) are shown as correlations. Where the scope of weighting is by production zone, weights are constant within each production zone, so the simple totals of production trip ends are replicated. This appears as a complete correlation in the table and a close replication of total observed trips.

Weighting by production zone:

- is efficient in that it reduces standard errors
- gives estimates similar to several other weighting schemes
- provides consistent weighting of zero and non-zero cells for each production zone
- guarantees the existence of observations where needed to calculate weights
- is probably closest to the original WTSM sampling and expansion scheme
- avoids some of the randomness in weights when calculated from few observations.

Although the expansion of the WTSM household survey is complicated, weights calculated by different methods or over different scopes often take the same value. Within a zone or wider sector, households with workers tend to have the same household expansion factor and workers tend to have the same person expansion factor. Even if there are multiple trips in a cell, they are all quite likely to be made by one worker or from one household. If there is no mixture of expansion factors for HBW trips within a weighting scope, the weight is simply the reciprocal of the expansion factor even if calculated as $\sum(\text{expansion factor})/\sum(\text{expansion factor})^2$.

The refinement of varied weighting is probably not vital for a well-conducted survey designed to achieve a consistent sampling rate. Variations in final weights may be a random effect, perhaps largely due to the sampling process and not requiring any further allowance. For simplicity and ease of checking totals while investigating other complex matters, a single weight was applied to all household survey observations throughout this study.

3.11.3 Other surveys – roadside interview surveys/external

Calibration of WTSM trip distribution and mode split included external movements surveyed by roadside interview, with a much higher sampling rate than the household interviews which provided the internal trips. Table 3.11 shows their respective expansion factors and, for ease of comparison, the inverse of their weights with allowance for uneven sampling.

Table 3.11 Expansion and weighting of household and roadside surveys

Scope	Count of observations	Trips (expanded)	Variance of trip total	Overall		
				Expansion	1/weight	1/weight
	Σ 1	Σ factor	Σ factor²	trips/count	var/trips	expansion
Internal	3045	183,215.72	13,692,579.4	60.2	74.7	1.24
External	1190	3901.29	15,696.6	3.28	4.02	1.23
All	4235	187,117.01	13,708,276.0	44.2	73.3	1.66

The inflation due to uneven sampling, shown in the last column, is very similar for the internal household and external roadside interviews. However, when the two surveys with their very different sampling rates are combined the inflation is much greater. Although including the roadside interviews with their much higher sample reduces the overall expansion rate considerably, the overall weight is hardly changed at all.

These large differences in sampling rates merit separate weighting scopes for the two surveys.

3.11.3.1 Systematic differences

When separate weighting scopes were applied, there were marked changes in the fitted coefficient of an Exponential model. In a correct model, weighting should not affect the fitted coefficients significantly; the correct weighting scheme should minimise the estimation errors. Different weighting schemes may vary the fitted coefficients within the limits of error by re-weighting the errors – in effect, reshuffling the same deck of cards.

Even if the model is not perfect, different weighting schemes may only re-arrange the errors (now including mis-specification) in the data at random. However, significant changes can occur when the weighting scheme is correlated with the errors. It was concluded that there were systematic differences in the fit of an Exponential model to the internal household and to the external roadside data. Possible causes are discussed in section 1.5.4 and some are analysed in section 4.12.

3.11.4 Scales

The effects of different scales of weighting are shown in table 3.12.

Table 3.12 Scales of weighting

Source	Effective expansion factor	Cost coefficient		Change in deviance	Mean residual deviance	
	1/weight	estimate	standard error	fitted	fitted	expected
Unweighted	1	0.06377	0.00018	336,449.8	21.600	0.758
WTSM estimate	40	0.06377	0.0011	8411.2	0.540	0.286
Simple expansion	60.17	0.06377	0.0014	5591.7	0.359	0.236
Complex expansion	74.73	0.06377	0.0015	4501.9	0.289	0.212
Person based	157.9	0.06377	0.0022	2130.7	0.137	0.140

At each scale, the same weight is applied to all observations, so the fitted coefficient is the same in all cases. This is the same coefficient as is shown in the bottom row of table 3.10 and the middle column of table 3.5.

Measures of fit and accuracy are scaled by the weighting. The weight reduces the expanded data to the effective sample size and hence the amount of information available. The fitted deviances are proportional to the weight and the standard errors are inversely proportional to its square root.

Expected deviance is not simply related to the weight, but to the sparsity of the effective sample.

Even with the unit weight, the expected deviance is less than unity. This indicates that a substantial part of the full 24-hour HBW car trip distribution is sparse even before sampling, with less than one trip expected (from an Exponential model) in many matrix cells. The fitted deviance is much larger than the expected value, showing how this weighting exaggerates the information in the data by ignoring its expansion from a smaller sample.

An overall sampling factor of 40 is given for the WTSM in TN17.1, section 2.5. This covers all purposes and modes, including roadside interviews and rail and school surveys which had higher sampling rates than the household interviews.

The simple weighting is the inverse of the overall expansion factors, $\sum 1/\sum$(expansion factor) summed over all records of HBW person trips by car from the household interviews.

The complex weighting, $\sum$(expansion factor)/$\sum$(expansion factor)^2, allows for the inefficiency of variable sampling rates. It was used for the original analysis of deterrence functions, but gave indications of exaggerated significance. The fitted deviance is still larger than expected, suggesting some systematic error or overdispersion in the model.

3.11.4.1 Units of observation

The preceding weightings based on the expansion of trips assume that observations of trips are independent. In a household interview survey (HIS) over a whole day, there are often multiple trips by the same person between the same production and attraction, usually from home to work and back for commuting trips. A more realistic approach is to take persons as the units of independent sampling, rather than trips, and to base the weighting on an expansion from persons, thus:

Person-based expansion factor = no. of trips made by person x trip expansion factor

The resulting overall weight of 1/157.9 includes the allowance for uneven sampling of the complex weighting, compared with a simple expansion factor of 105.3 from 1740 observed commuters to 183,215.7 trips. The differing number of trips per person contributes to this variation in sampling rate.

The final row of table 3.12 shows that, with this weighting, the fitted deviance is slightly less than expected, and there is no longer any suggestion of residual systematic error or overdispersion. Table 4.21 shows similar results for other deterrence functions, and the changes in deviance from over-fitting shown in table 4.24 are also consistent with this weighting. It has been adopted throughout this study.

Since the weighting is based on an expansion from persons as units of observation, it is most readily calculated from a dataset of persons. Alternatively, if it is calculated from a dataset of trips, each trip should be scaled down by the number of trips made by that person, otherwise there is bias towards the greater number of trips made by the more active trip makers.

3.11.4.2 Land-use formulation

Person-based weighting evolved from pairing the home and workplace for each worker and treating each such pairing as one observation. This reformulation is described in appendix B; it was inspired by proposals to derive the Auckland trip distribution from a land-use model, rather than a transportation

model. When these proposals were not adopted, it was decided to continue working with the WTSM trip-based formulation, but adopt the weighting suggested by the pairings of homes and workplaces.

A comparison of trip-based and land-use formulations with common datasets is given in table 3.13 for calibrations of a simple Exponential trip distribution.

Table 3.13 Calibration from trip-based and land-use formulations

	Expanded trips (WTSM)		Home-workplace pairings	
	All	Common		All
Observed persons	1740	1621		1969
Observed trips	3045	2737	~	~
Expanded trips	183216	164,609.3	~	~
Weight	Trips/157.9		Unity per person	
Fitted coefficient	0.0638	0.0617	0.0585	0.0576
Standard error	0.0022	0.0023	0.0018	0.0016
Change in deviance	2130.7	1868.6	2747.3	3253.3
Mean residual deviance	0.1368	0.1261	0.1868	0.2026
Expected mean residual deviance	0.1399	0.1347	0.1788	0.1926

Compared with the WTSM trip-based formulation, the common dataset excludes:

- HBW trips where the attraction meshblock is not that of the identified workplace
- persons who do not make a weekday HBW trip by car to their workplace
- persons who are not coded as workers in the household survey – possibly people sharing a car for a different purpose
- visitors to the household.

The set of all home-workplace pairings includes all workers in the common dataset and those who:

- make any weekday HBW trip by car, or
- do not make any weekday HBW trip by another mode.

This mainly adds workers who do not make direct trips between home and work. Their primary mode may not be car, though it is usually most convenient for such trip chaining.

The fitted coefficients are reasonably close in practical terms. The difference between the two formulations of the common datset is on the same scale as the standard errors, implying that differences in definition may be as large as those arising from sampling.

Standard errors are larger and changes in deviance are smaller for the common datasets than for the full datasets for the same formulation, roughly in accordance with the reduction in sample size. The ratio between the changes in deviance on the common dataset, 1868.6/2747.3 = 0.68, is very similar to the ratio between the simple expansion factor from observed persons to expanded trips, 105.3, and the weighting factor including allowance for varied sampling and trip rates, 157.9 (105.3/157.9 = 0.67). The simple weighting of observed home-workplace pairings omits variations in both trip rates and sampling rates which occur in the expansion to trips.

Mean residual deviances are all consistent with their expectations and do not show the overdispersion apparent with other scales of weighting in table 3.12.

The fundamental mechanism of trip distribution is the choice of home and workplace, a long-term decision. In the trip-based formulation, the weighting of home-based work trips is heavily influenced by the occurrence of intermediate calls between home and work, possibly the most transient of decisions.

The assumption in the land-use formulation of one workplace per worker involves a degree of approximation which may be reasonable in the context of broad land-use/transport modelling. It may also be reasonable for the purpose of education – one school per child. It becomes implausible at the household level – one workplace or school per household – and ridiculous for other purposes such as shopping – one shop per household (except in a command economy, where the transport problem of linear programming can be applied to simply minimise cost).

These issues in the systematic modelling of travel may be addressed better in models based on tours or activities. Any sensible error model needs to acknowledge that the origin of a trip is not independent of the choice of destination of the previous trip; this is addressed by person-based weighting in this study.

3.11.4.3 Tour- and activity-based modelling

The land-use formulation is similar to tour- or activity-based modelling in that it identifies a primary attraction for a sojourn or chain of trips away from home which may include stops at other attractors before returning home. It may go beyond these advanced formulations for modelling travel in that it omits the rate of trip making from the calibration, save for excluding those who do not travel to work.

Identifying trip chains for tour or activity modelling is complicated, involving many of the issues encountered in the land-use formulation (appendix B). This research has adopted the trip-based dataset provided by the WTSM, but its methods of calibration are also applicable to tour- or activity-based modelling, or potentially any model of travel which is deterred by cost.

3.11.5 Level of analysis – sampled or expanded trips, with offset or weighting

Sen and Smith (1995, section 5.9.3) describes two methods for representing the scale of errors in an expanded sample, by offset and by weighting.

A trip distribution model with an Exponential deterrence function can be written as

$$\text{expanded trips} = P_i A_j \exp(-\lambda C_{ij})$$

where P_i and A_j combine trip end totals and balancing factors. This can be rewritten as a log-linear model for fitting as a GLM

$$\log(\text{expanded trips}) = \log(P_i) + \log(A_j) - \lambda C_{ij}$$

The expanded trips are the dependent (Y) variable with a logarithmic link to a linear model of dependent (X) variables, comprising dummy variables or factors for the production and attraction zones, and the cost. Substituting

$$\text{expanded trips} = \text{observed count} \times \text{expansion factor}$$

gives

$$\log(\text{observed count}) + \log(\text{expansion factor}) = \log(P_i) + \log(A_j) - \lambda C_{ij}$$

$$\log(\text{observed count}) = -\log(\text{expansion factor}) + \log(P_i) + \log(A_j) - \lambda C_{ij}$$

The observed count is now the dependent variable and the linear model includes the term log(expansion factor). Unlike the other terms, no coefficient is to be fitted; such a term is introduced into a GLM as an 'offset'. It reduces the other linear terms, which constitute a distribution model of expanded trips, back to

the scale of observed counts. Since the observed count is being modelled, the Poisson errors expected in a count are appropriately scaled too.

The alternative approach to scaling the errors is to enter weights as a weight term in the GLM. In this case the systematic trip distribution is modelled on the scale of expanded trips, but the errors are changed to the scale expected from sampling by the weight term.

The two approaches can give identical results for simple cases where weights are the reciprocal of the expansion factor. As has been shown above, this is not the case for variable sampling rates, whose inefficiency reduces the effect on accuracy of the observed count. Models using an offset could be fitted to a reduced 'effective observed count', but this would add complexity. The fitted values from an offset model have to be re-expanded to represent the full population of trips. For these reasons the weighting approach has been adopted exclusively in this study, after checking the consistency of the two methods in the simple case.

3.12 Summary

The cost coefficient for a trip distribution deterrence function may be calculated from a 'four-square' set of data with just two production zones and two attraction zones. The calculation shows the cost coefficient relates interaction effects in trip making with non-additive differentials in costs.

In practice, matrices are much larger, leading to redundancy in information, but observations are subject to error. Coefficients can be estimated by maximum likelihood using generalised linear models (GLMs); the log-linear form is consistent with the multiplicative form of the distribution model and a Poisson sampling error.

Fitting such models replicates total trips and trip costs for their simplest form of deterrence function, the Exponential. Where they fit K or L factors, they replicate trips and trip costs respectively over the scope of those factors. In particular, trip ends are replicated by the trip end balancing factors which act as zonal K factors.

There are well-established statistical measures for the fit of GLMs. In general, these are approximations to the measures for simple regression or analysis of variance with Normal errors, with the deviance ($-2 \times$ log-likelihood) similar to squared residuals. However, observed trip matrices are often sparse in practice, with less than one observation per cell. The general expectation that a well-fitting log-linear model with Poisson-like errors will have a mean residual deviance of unity is no longer true under sparsity, and Pearson's chi-square, an alternative statistic, becomes unreliable because of increasing variance. However, the change in deviance between nested models is relatively robust under sparsity, so it can be used to test the significance of variables as they are added to a model, even if the goodness-of-fit of any particular model is hard to judge.

Analytical approximations have been derived for the expected mean and variance of the residual deviance of well-fitting log-linear but sparse models with Poisson-like error. When sparse data is disaggregated, the relative error in the residual deviance appears to depend upon the total count of observations in the sample rather than the number of records or matrix cells over which it is disaggregated.

The change in deviance when a relevant variable is added in to a model can be seen as reflecting the amount of information about the effect of the variable in the dataset. This information is reduced when data is aggregated across records or cells if independent model variables differ, and are averaged in aggregation.

Because the residual deviance does not provide a reliable scale of overall error, the model has to be weighted to reflect sampling effects as the primary source of error. Data that has been expanded from a sample count to represent a whole population has to be weighted by the sampling fraction to represent the effect of Poisson error arising from sampling.

GLMs allow separate weights by individual record. This may be used to represent different sampling rates in different surveys. If each observed trip has its own expansion factor, it can have its own weight calculated from it. However, this cannot be used to weight matrix cells with zero observations, so weights have to be calculated over scopes that include such cells. Where there is a mixture of sampling rates, the overall weight is not the simple ratio of total observed counts to total expanded trips because of the inefficiency of unequal sampling.

While the value of a single, common weight applied to all records affects statistical measures of significance and accuracy, it does not affect the estimate of the cost coefficient, or of other model coefficients or the fitted trip distribution. Where the weighting of records varies, the fitted coefficients and trip distribution can change according to the relative weighting. Log-linear Poisson models fitted with differential weighting replicate the **weighted** average of each model variable, complicating interpretation. A single common weight for the WTSM HIS gave results similar to more complex weighting schemes, and was therefore adopted for simplicity. Weighting schemes need to be designed with the survey sampling and expansion schemes.

Zero cells represent movements which could be observed by the survey, but the observed count was zero. They represent information that the movement is small. They should be included in analysis and properly weighted to reflect the probability of observing such a movement from the sampling. Their omission or zero-weighting is shown to severely distort results.

Null cells represent movements which could not be observed by the survey. They contain no information about the movement, and should not be included in the data for calibration. There are no null cells in the fully observed internal trip matrix from the WTSM HIS used in this study. Null cells commonly arise from roadside interview surveys, and their estimation by a GLM is equivalent to partial matrix infilling methods.

Empty zones are whole production or attraction zones which have no observed trip ends, appearing as complete rows or columns of zeroes in the trip matrix. Although they may have been observable from the survey, they contain no information about trip distribution. They are liable to complicate computation and dilute the mean residual deviance, and are better omitted from the calibration of trip distributions. Matrix cells are empty by virtue of representing a movement to or from an empty zone.

In conventional transport models, the unit of observation is the person-trip, from origin to destination. However, trips by the same worker are likely to be to and from the same workplace, and the production-attraction pairings of such trips will not be independent. Therefore the weighting was reduced to reflect workers as the independent units of sampling, rather than work trips.

GLMs can be fitted either to the original sampled counts, or to the trips expanded from them. Where there is a single, common expansion and hence weighting factor, the two approaches can be made equivalent by simple adjustment of the offsets or weights. With mixed expansion factors, and weighting adjusted for consequent inefficiency and for workers as the independent unit of observation, it is easier to work by weighting the expanded trips.

4 Deterrence functions

4.1 Introduction

This is an exploration of alternative forms of deterrence function used in trip distributions. Deterrence functions describe how the probability of choosing a destination varies with the cost of the journey. For a general description of deterrence functions, see section 2.3.

The Exponential is taken as the base deterrence function, with other forms developed from it. It is:

- consistent with theories of random utility and maximum entropy
- mathematically tractable and the simplest form of log-linear model
- adopted in the WTSM.

Deterrence functions are plotted on the log scale; consequently the Exponential function appears as a straight line. Because they are relative functions, the vertical origin is arbitrary.

4.1.1 Trips

Trips are taken from the WTSM household interview survey (HIS). They are 24-hour weekday production-attraction, home-based work (HBW) person trips by private vehicle.

They include all trips from the WTSM choice car household/mode segment, most from the competition car/slow segment (excluding slow modes) and a few from the captive segment (possibly passengers). Only internal trips are analysed. There are 3045 trips by 1740 workers counted in the sample, which are expanded to 183,216 daily trips.

Empty trip ends are excluded, leaving 162 production zones and 194 attraction zones, giving 31,428 matrix cells.

4.1.2 Costs

Costs are taken from the distribution synthesis stage of the base WTSM model. They are generalised costs, with components of time and distance (and potentially tolls) in units of minutes. The composition is specific to trip purpose, but applies across periods because trips are distributed in all-day matrices. Factors used are:

- vehicle operating costs = 15 cents per km
- value of time = 13.6 cents per minute (car owning household, for HBW)
- 24 hour = 0.565 AM + 0.435 IP (no PM assignment).

Parking charges are applied to trips with attractions in central Wellington. The values of

- $1.70 – lower Wellington
- $2.75 – upper Wellington

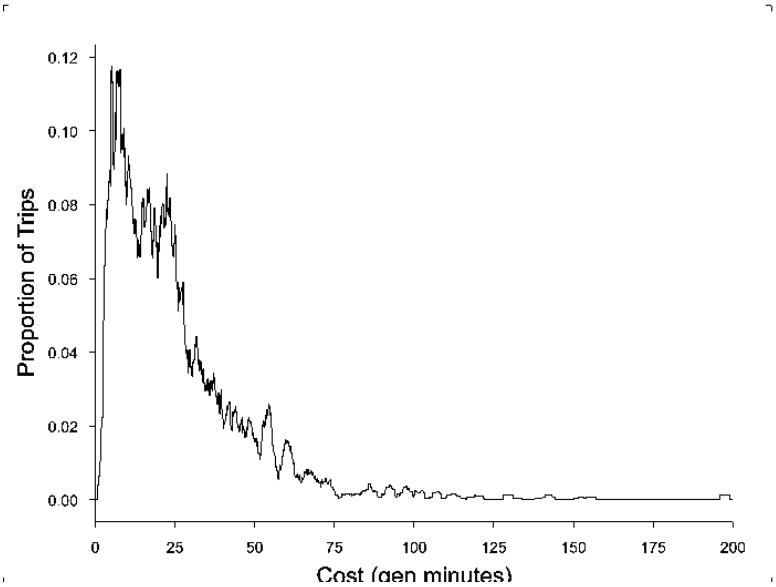
are halved, to divide a single parking fee between outward and return trips. Because the charges are applied across all attractions to a zone, they do not affect distribution with an Exponential deterrence function, but they do affect mode split.

Vehicle operating costs and parking charges are divided by the vehicle occupancy of 1.19 for HBW to give costs per person trip.

There is an iterative distribution–assignment loop. The generalised costs are subject to damping by averaging with previous values.

The distribution by these costs of observed trips is shown in figure 4.1.

Figure 4.1 Cost distribution



This distribution is also shown by geographic segment in figure 4.21, and cumulatively in figure 4.9.

Table 4.1 Ranges of costs (generalised minutes)

Application	Minimum	Maximum
Calibration – between non-empty zones	0.61 (1.21 interzonal)	276
Calibration – cells with observations	0.75 (1.21 interzonal)	196 (162 external)
Synthesis – all cells	0.41 (0.82 interzonal)	276 (288 external)

Table 4.1 shows ranges of costs. Minimum costs are all intrazonal, so the minimum interzonal costs are also given. External movements are excluded from these analyses, but they do not greatly affect the range. Limits on other scales are given in table 4.5. Upper limits are plotted in figure 4.36. These minima and maxima are described in more detail in appendix A, with an example of their calculation.

4.1.3 Weighting

All weighting is by a single factor of 1/157.9, based on the expansion from HBW trip-making persons in the sample to the population of all weekday trips. Treating persons (and hence their workplaces) as independent observations allows for the lack of independence between the attractions for most commuter trips made by any individual person.

Earlier analyses based on the expansion from observed trips, with a single weight of 1/74.73, gave the appearance of significance in the improvements with more complex models, in particular in higher orders of splines and polynomials, hence some fitting and analysis beyond their current significance.

4.1.4 Measure of statistical fit

The principal measure of fit used to compare models is the deviance, a statistical measure of likelihood. It is similar to the sum of squares in simple regression. For a Poisson error model

$$\text{deviance} = -2(\text{observed} \times \log(\text{observed}/\text{modelled}) + (\text{observed}-\text{modelled}))$$

Deviance is minimised in fitting a GLM; the lower the residual deviance, the better the model fits.

Statistical fit is shown graphically, with the residual deviance plotted against the residual degrees of freedom. Residual degrees of freedom are reduced as more terms are introduced into the model, or the order of a spline or polynomial curve is increased. Basic models are plotted on the left; models increase in complexity with more terms towards the right, so the residual degrees of freedom decrease in this direction.

The effect of introducing a variable can be seen by joining equivalent models with and without the variable. At worst, the line would be flat, showing the variable did not improve the fit at all. Even an uncorrelated variable would be expected to improve the fit slightly, giving a very slight slope. The steeper the downward line, the higher the significance of the variable.

Changes in deviance are tested against the χ^2 distribution, as advised by Payne et al (2009, p304). Significance is at the 5% level, unless stated otherwise.

4.1.5 Structure of this chapter

The next five sections consider different forms of deterrence function.

Section 4.2 deals with basic analytical forms. The Exponential is a natural form in GLMs. Its relationship to the Power function by logarithmic transform is demonstrated. The combination of the Exponential and Power forms in the Tanner function is considered as another transform of costs, akin to generalised costs with time and distance components.

Empirical models, based on cost bands, are introduced in section 4.3. GLMs can calibrate more sophisticated forms than the classic flat-topped step. Their parameterisation is described and their effects shown on cumulative residual plots leading to suggestions on the choice of break points between cost bands.

The WTSM adopted a sophisticated geographic segmentation. This is an example of fitting K factors or constants to different segments of a matrix. In the WTSM different cost coefficients or L factors are also fitted. The WTSM formulation is described in section 4.4 and re-fitted by GLM for comparison with other forms.

Splines are a continuous form of empirical curve which can be fitted by GLMs. Section 4.5 describes their fit up to high orders, where improvements in fit are small. Some parts of the function can become counter intuitive, showing preference for more distant destinations, so their turning points are considered.

Polynomials are a more traditional method of exploring curvilinearity, with analytical forms. They raise the same issues as splines and are described in section 4.6.

Non-linear functions such as the Box-Cox which cannot be fitted by basic GLMs have been calibrated by extensions to the algorithms in section 4.7.

Section 4.8 brings together these different forms of deterrence function to examine their statistical fit measured by deviance. The residual deviances and insignificant changes in them are compared with their expectations. The extent to which different deterrence functions account for the systematic changes in deviance is discussed next, leading to consideration of sample size.

Section 4.9 examines more practical measures of the fit of trip distributions. Screenline crossings can be compared with independent traffic counts, while predicting scheme usage and benefits is the raison d'être of

transport models. Differences between deterrence functions are compared with differences between observed matrixes and traffic counts on screenlines, and between different screenline and scheme locations.

The various forms of deterrence functions are all calibrated on a common set of trips and costs, described in sections 4.1.1 and 4.1.2. Section 4.10 considers other definitions of cost, starting with crude crow-fly distances and moving on to re-examine the factors relating the components of generalised cost, which are time and distance from the morning and interpeak periods. Sensitivity to the formulation of intrazonal costs is considered in section 4.11 and alternative costs to external zones are examined in section 4.12.

Section 4.13 summarises the findings.

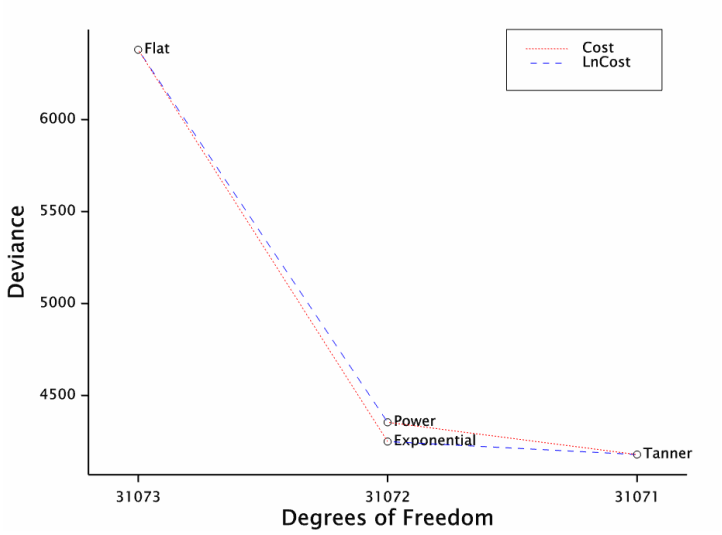
4.2 Analytical functions

These are continuous curves which are well established as deterrence functions. The Exponential is supported by maximum entropy and choice theories, reproduces observed travel costs, and is simply fitted in a GLM. The Power function is fitted by substituting the logarithm of cost for cost; it gives a constant elasticity, and is the same form as the inverse square law of the physical gravity model. The Tanner function, also known as the gamma function, combines the two. See section 2.2 for further discussion.

4.2.1 Fit

Figure 4.2 shows the fit of these models, starting from the flat model. The flat model accounts for variations between zones in production and attraction trip ends; trips are simply proportioned by these observed trip end totals, without any effect of cost. This is the base model for all deterrence functions, but is omitted from later plots of deviance to show differences between deterrence functions more clearly.

Figure 4.2 Fit of analytical functions



Both cost and log(cost) improve the fit of the flat model greatly; the deviance is reduced by about 1/3, for a reduction of 1/31073 in degrees of freedom. The Exponential model with cost fits better than the Power model with log(cost). The improvements in adding the other term, to produce the Tanner model, are much smaller since the two terms are highly correlated, but still very significant.

4.2.2 Fitted models

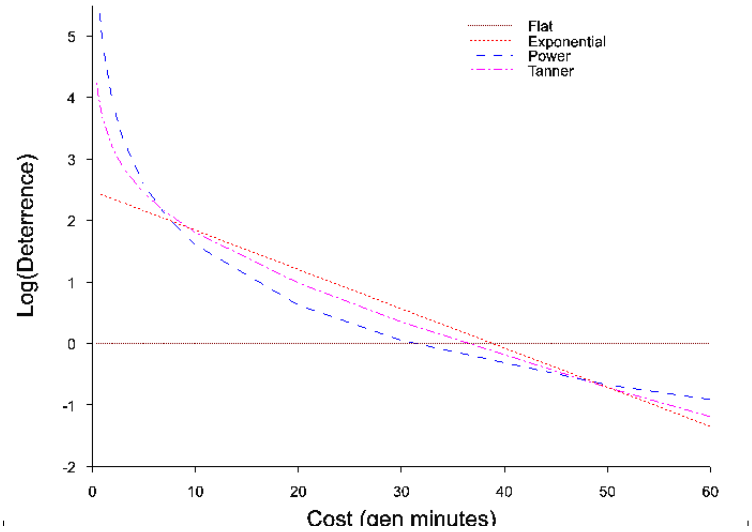
Table 4.2 Fitted coefficients of analytical functions

Model	Cost			Log(cost)		
	Coefficient λ	Standard error	t ratio	Coefficient γ	Standard error	t ratio
Exponential	0.06377	0.0022	28.7			
Power				1.416	0.033	43.3
Tanner	0.0364	0.0035	10.4	0.662	0.074	9.0

Although the Exponential gives the better fit according to the deviance, the Power coefficient has a higher t ratio.

Combined in the Tanner, coefficients are reduced and standard errors are increased, giving much reduced but still highly significant t ratios. This reduction in standard errors is an indicator of the high correlation of the two coefficients, which is –0.836.

Figure 4.3 Deterrence functions – analytical



The fitted deterrence curves are plotted in figure 4.3. The deterrence is plotted on a logarithmic scale, with greater probability of trips to the top. The location of curves on this vertical axis is arbitrary, since differences are absorbed in the trip end balancing factors.

With these scales, Exponential functions appear as straight lines, while the flat model lives up to its name. The fitted Power and Tanner models take concave forms, with short and long trips relatively more likely than medium ones. The Tanner function is less concave than the Power. It does not have a turning point and decreases with very low costs, as its form allows, because the coefficient of log(cost), γ, is positive as well as the cost coefficient λ (table 4.2). All functions decrease monotonically with increasing positive costs.

Figure 4.4 Deterrence functions – analytical, on log scale

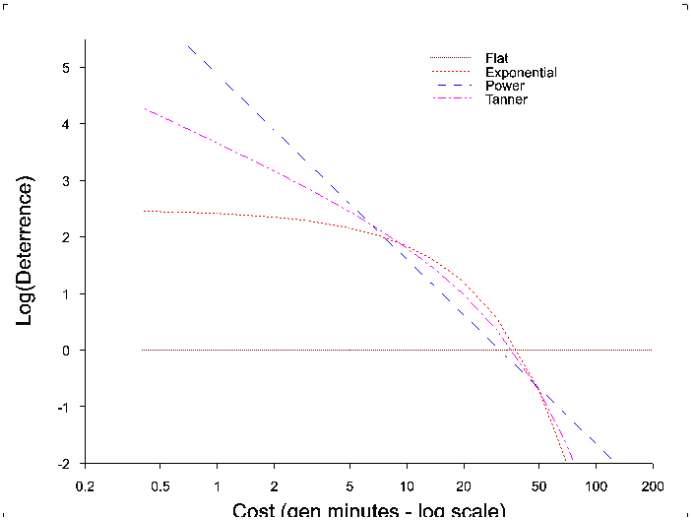


Figure 4.4 plots the deterrence functions against the logarithm of cost. On this scale, the Power function is a straight line, reflecting its constant elasticity. The better fitting Exponential and Tanner functions now appear convex.

Figure 4.5 Residuals – analytical functions

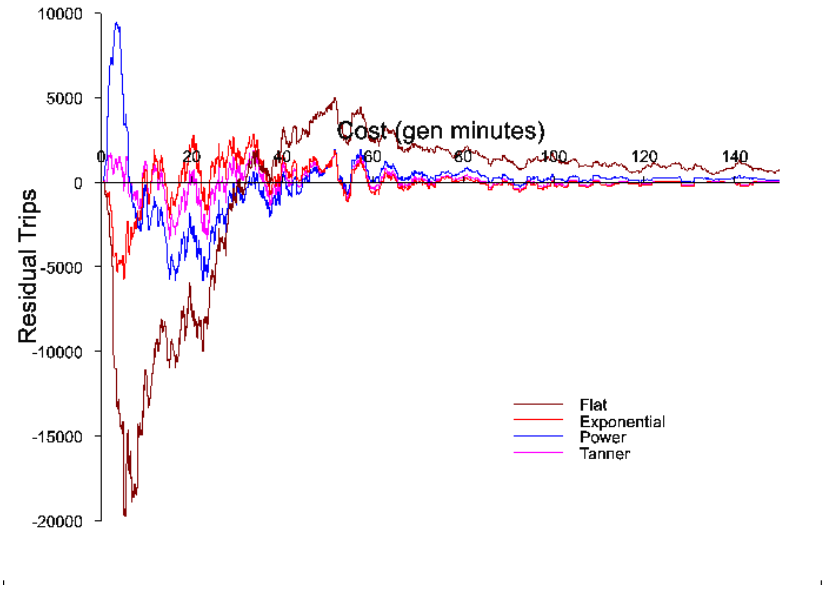
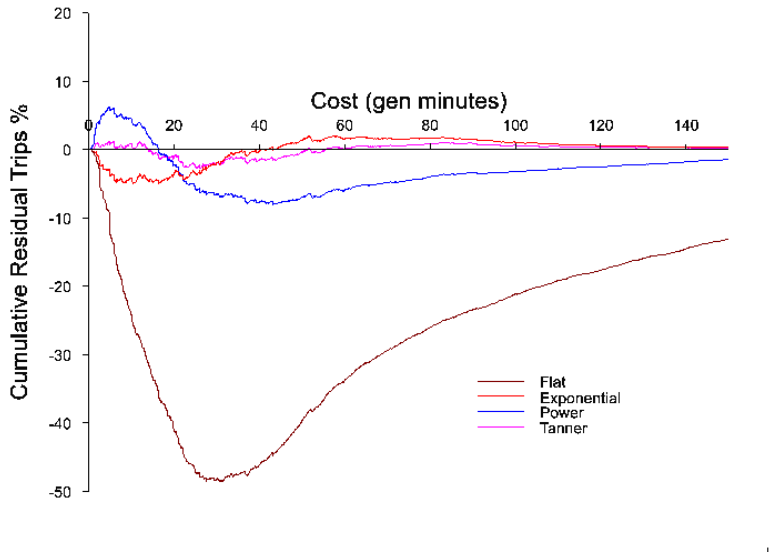


Figure 4.5 shows the residuals of modelled trips after subtracting observations. Because the flat model has no sense of cost, it underestimates short trips and overestimates long ones. The Exponential struggles to estimate enough very short trips, but the Power overdoes it amongst the shortest trips, which are typically intrazonal (see figure 4.21).

These residual plots above are the difference between modelled values and the observations plotted in figure 4.21; the differences are small and hard to see among the noise if plotted in that figure. Some aspects are easier to see in terms of cumulative residuals, as in figure 4.6. These are the difference

between modelled values and the observations plotted in figure 4.9; again, the differences are small when plotted on that figure.

Figure 4.6 Cumulative residuals – analytical functions



The areas between the curves and the horizontal axis are the differences in total travel cost. Because the Exponential model reproduces the observed costs, its areas above and below the axis balance out. Since the Power model reproduces log(cost) its areas would balance with a logarithmic horizontal axis. Although this balance does not appear on the natural scale above, it is still a constraint tying the curve to the axis. The Tanner model, subject to both constraints, follows the axis more closely and appears to give a better overall match for the whole range of generalised costs. The flat model was never in the game.

Table 4.3 shows how closely these total travel costs are reproduced. All models reproduce total trips with a few parts per million computational error. Similar accuracies are achieved for costs with corresponding terms in the model – cost in the Exponential, log(cost) in the Power and both in the Tanner.

Table 4.3 Fit of trip and cost totals by analytical functions

Model	Trips	% difference	Cost	% difference	Log(cost)	% difference
Observed	**183,216**	~	**4,451,140**	~	**513,167**	~
Flat	183,231	0.0082	*12,707,921*	*185.5*	*718,649*	*40.0*
Exponential	183,217	0.0005	4,451,161	0.0005	*528,856*	*3.1*
Power	183,219	0.0018	*5,453,493*	*22.5*	513,173	0.0013
Tanner	183,217	0.0007	4,451,165	0.0006	513,170	0.0006

Bold – observed data, base for differences

Italic – totals without corresponding terms in the model

In the flat model, without any cost terms, the error in log(cost) is smaller than in cost, probably because the logarithmic transform reduces the range of longer movements with reduced trip making that the flat model fails to recognise. The error in log(cost) in the Exponential, 3.1%, is much smaller than the error in cost in the Power, 22.5%, even relative to the errors in the flat model. This again suggests that the Exponential is a better model than the Power for this data.

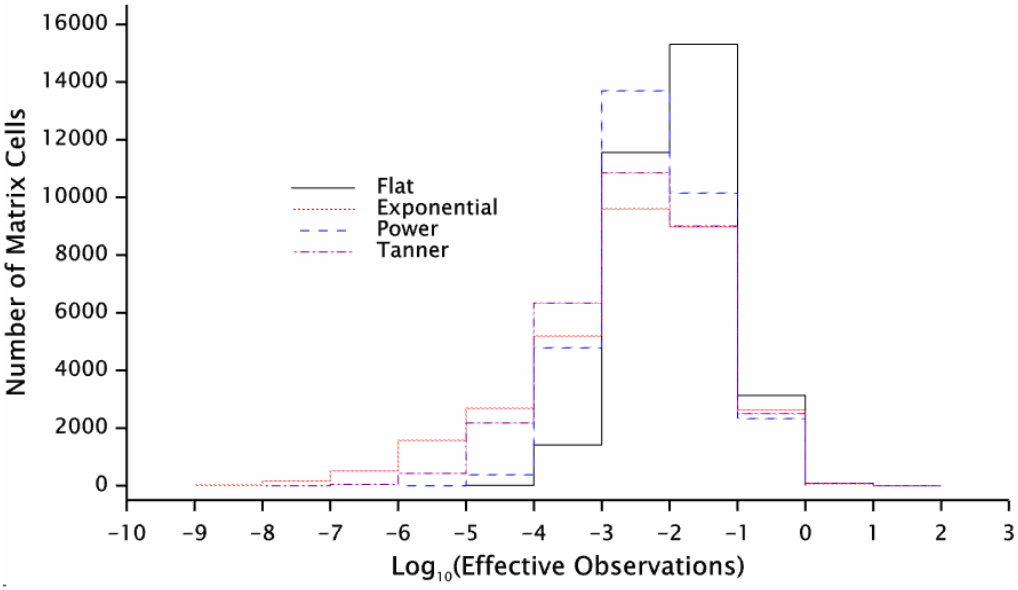
The better of the models considered in the following sections tend to reduce errors in the totals of natural or logarithmic costs to 1% and less, even if they do not include a corresponding term. Computational errors where a term is fitted can increase in high order splines and polynomials, but are still typically 0.1% or less.

4.2.3 Sparsity and Pearson's chi-square statistic

These analytical models show marked differences in two ways: the scale of predicted cell values and Pearson's chi-square statistic.

Figure 4.7 shows the distribution of cell values fitted by the different models. The horizontal axis is a logarithmic scale (10^{-10} to 10^{+3}) of effective observations, ie the modelled values weighted by 1/157.9, approximately the sampling factor of workplaces. Statistical measures such as deviance are calculated on this scale of observed counts, rather than expanded trips. Zeros are excluded; they occur only for empty zones where no trip ends are observed either at the production or at the attraction zone.

Figure 4.7 Distribution of fitted values



All distributions have an arithmetic mean of 0.037 or $10^{-1.43}$ since they all reproduce trip numbers; this is sparse in terms of Poisson deviances. The flat model has a relatively narrow spread of cell values, since it has no cost deterrence. All cost models have similar numbers of cells with large values on the right, but the Exponential has many more cells with very small values on the left of the plot. This is to be expected since the fitted Power and Tanner functions curve above the Exponential at high costs.

The minimum cell value in the Exponential model is 0.22×10^{-6} trips, or weighting by 1/157.9, 1.4×10^{-9} effective observations. Most models considered in the following sections have minima that are one or two orders of magnitude larger, although high-order natural splines and polynomials can be even smaller.

Pearson's chi-square statistic is an alternative measure of fit to the deviance.

$$\text{Pearson's chi-square} = (\text{observed-modelled})^2/\text{modelled}$$

Like the deviance, it approximates to the normalised sum of squares when modelled values are not small (>1), and is identical in ordinary regression with normal errors. Unlike the deviance it is not minimised in fitting a GLM, so it does not necessarily decrease as more terms are added into a model. This would still

be expected, but it is dramatically greater for the Exponential model than for the flat model. It is lower only for the Power model.

Table 4.4 Residual statistics of analytical functions

Model	Pearson's chi-square	Deviance	Degrees of freedom
Flat	33,040	6380.4	31,073
Exponential	1,154,403	4249.7	31,072
Power	22,500	4354.5	31,072
Tanner	60,487	4179.2	31,071

This is probably an artefact of the sparsity shown in figure 4.7. The deviance and Pearson's chi-square both have expected means of unity when the data is not sparse. While the expected deviance decreases with sparsity (figure 3.4), the expectation of Pearson's chi-square stays at unity. However, the variance of the Pearson statistic becomes very large with sparsity, while that of the deviance decreases.

It was concluded that Pearson's chi-square is not a satisfactory test statistic for this analysis.

4.2.4 Tannerised cost

The Tanner deterrence terms

$$\lambda \times \text{cost} + \gamma \times \log(\text{cost})$$

can be re-written as

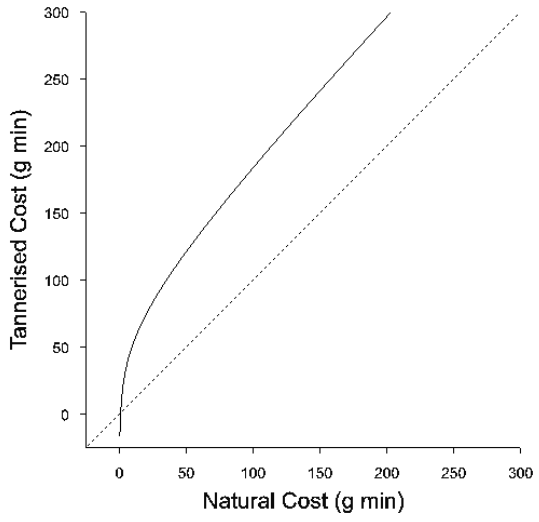
$$\lambda \times (\text{cost} + \gamma/\lambda \times \log(\text{cost}))$$

The bracketed term has the dimensions of cost, and is akin to the generalisation of cost by the inclusion of time and distance components. This term is therefore referred to here as the Tannerised cost. Although the two components of the coefficient γ/λ are highly correlated, it has a reasonably small standard error, though the t ratio is again smaller than that of its component coefficients.

γ/λ = 18.20, standard error 1.32, t ratio 13.8

The relationship this coefficient gives between Tannerised cost and cost is shown in figure 4.8.

Figure 4.8 Tannerised cost



The dimensions of both axes are minutes, since cost itself is being measured as generalised time. The Tanner cost is negative for natural costs just below one minute, and the steep slope in that region may be related to intrazonal trips. The range of costs is shown in table 4.5.

Table 4.5 Extent of cost distributions

Measure of cost distribution	Natural	Logarithm	Tanner
Lowest cell	0.4096	-0.8926	-15.84
Lowest of non-empty cells	0.6066	-0.4999	-8.493
Lowest cell with observed trips	0.7479	-0.2905	-4.540
Mean trip cost	24.29	2.801	75.28
Highest cell with observed trips	196.0	5.278	292.1
Highest cell	276.0	5.620	378.3

Units: minutes of generalised time

Fitting the Tannerised cost in a simple Exponential model reproduces the cost coefficient λ from the Tanner model (0.0364) with the same residual deviance. However, the standard error is much reduced at 0.0012, giving a t ratio of 30.8. This reflects the greater certainty after fixing the correlated coefficient of the log term. Both standard error and t ratio are better than for the original Exponential, but the t ratio is still not as high as the Power model, with its worse fit according to residual deviance.

In practical terms, Tannerisation causes distributions to be affected by trip end costs common to all movements, such as CBD parking charges and the distance of external zones from their entry points. These are neutral in a simple Exponential deterrence function. The same issues arise with the Power function, whose logarithmic transform is a special case of Tannerisation.

Tannerised cost might be used for comparing other deterrence functions with the Tanner as a base. It avoids the profligate fitting of a separate coefficient of log(cost) as well as cost to every segment in a function. This is at the expense of not fitting any such coefficient within the model, but taking it 'fixed' from the model discussed above. The simultaneous estimation of a common mix of logarithmic and natural cost in a multi-segment model falls outside the linear form of GLMs. A basic Tanner model is fitted in this way as a non-linear function in section 4.7.

Geographic segmentation models with Tannerised costs have residual deviances similar to those of geographic segmentation models with a log(cost) term added, shown at the bottom of figure 4.30. This was also found with the five-slope empirical model.

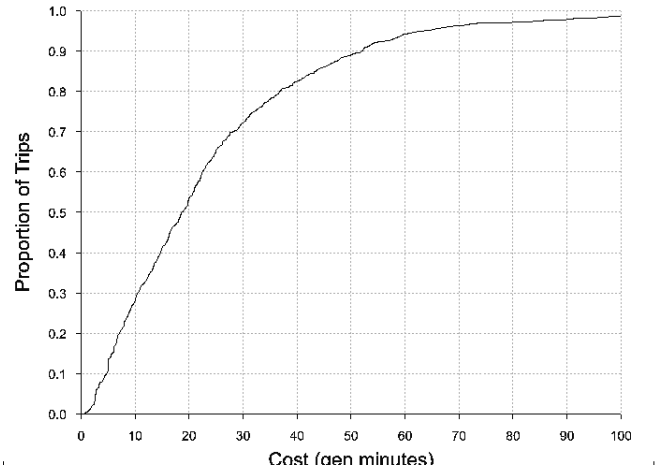
Mixtures of natural and logarithmic costs that combine to include a Tanner function were tried in formulations of splines (section 4.5) and polynomials (section 4.6). Splines and polynomials of Tannerised costs are broadly similar.

4.3 Empirical functions

4.3.1 Form and parameterisation

The empirical approach is to divide cost into ranges or bands, and fit a separate deterrence value to each band without any analytical relationship between them.

Figure 4.9 Cumulative cost distribution



Inspection of figure 4.9 suggests three sets of convenient breakpoints for defining cost bands:

Number of bands	Break points
2 – crudest possible	**20**
5	10, **20**, 30, 50
10	*5*,10,*15*,**20**,*25*,30,*40*,50,*60*

The breakpoints of the smaller sets also appear the larger ones. The models are thus nested: a model with more bands is a direct development of one with fewer bands, and incorporates it.

The breakpoints have been chosen to give roughly the same number of observed trips in each band, ie dividing the vertical axis equally when projected on the cumulative curve. Upper breakpoints have been chosen to divide broad cost ranges even if there are relatively few observed trips. Despite this, upper bands still contain greater proportions of travel costs. Table 4.6 shows the distribution of observed trips and costs between the bands.

Table 4.6 Observed trips and costs – by empirical band

Band	Trips				Travel cost (generalised minutes)				
(gen min)	2 band	5 band	10 band		2 band	5 band	10 band	Total	Mean
0–5	53%	28%	13%	24,683	22%	6%	2%	84,146	3.41
5–10			14%	26,389			4%	196,802	7.46
10–15		25%	13%	23,997		16%	7%	299,307	12.47
15–20			12%	22,392			9%	393,180	17.56
20–25	47%	19%	12%	21,932	78%	19%	11%	491,230	22.40
25–30			7%	12,904			8%	352,335	27.30
30–40		17%	10%	18,906		27%	15%	653,053	34.54
40–50			6%	11,864			12%	530,367	44.70
50–60		11%	5%	9519		33%	12%	520,053	54.63
60 +			6%	10,629			21%	930,668	87.56
Overall	**100%**			**183,216**	**100%**			**4,451,140**	**24.29**

The break points are specified as upper bounds, so all intrazonal trips fall in the lowest band. Even when this is limited to five generalised minutes, some 30% of trips in it are still between zones, rather than intrazonal.

4.3.1.1 Flat steps

The classic model has a single value of deterrence across the whole of each band. This gives a stepped form when plotted against cost. The categorisation allows it to be treated as a third dimension of a matrix and fitted using iterative balancing techniques of Furness or Fratar.

In Genstat, this categorisation is conveniently handled by converting costs to a 'factor', a set of dummy (0,1) variables representing the band that each cost falls into. Factors can be treated as a single variable, with many of their properties recognised and handled internally by the software. Similar facilities are available in other statistical packages.

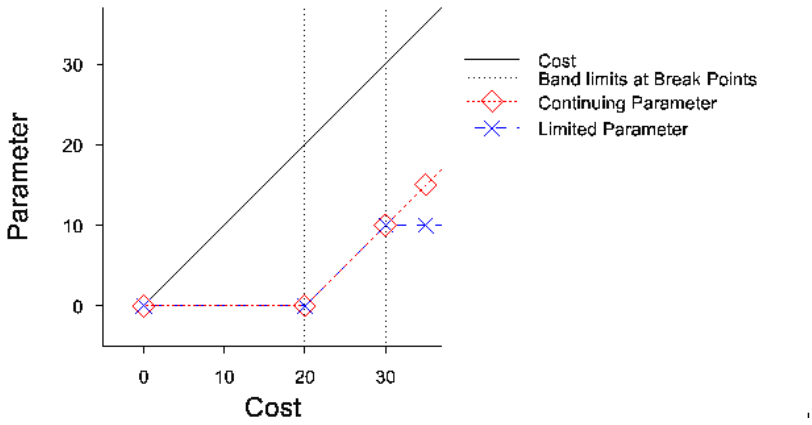
4.3.1.2 Slopes

GLMs can fit a wider variety of models within the banded structure. Slopes can be fitted within each band. The natural form in a log-linear model is the Exponential, appearing as a straight line on log(deterrence) plots.

A common slope can be fitted across all bands, leaving steps between the bands. Alternatively, the slopes can be formulated to differ in each band, but join without steps.

Technically, this is achieved by converting the costs into a set of variables, one for each band, akin to a factor. The value of the variable for a band is zero below the band's lower limit, and then increases with cost from that limit. The variable may continue to increase with cost above the upper limit of the band, or remain constant at its value for the upper limit, as shown in figure 4.10. These give different parameterisations of the same model. In the first case, the fitted coefficients are the difference from the previous slope; in the second they are the absolute value. The different sets of standard errors can help interpret the fit from different perspectives.

Figure 4.10 Parameterisation of L factor slopes



Example only (values are for the third of five bands)

This technique for slopes is sometimes known as 'broken stick'. Factors can similarly be re-parameterised into a set of step functions, but this loses some convenience of handling a factor as a single entity in the software package.

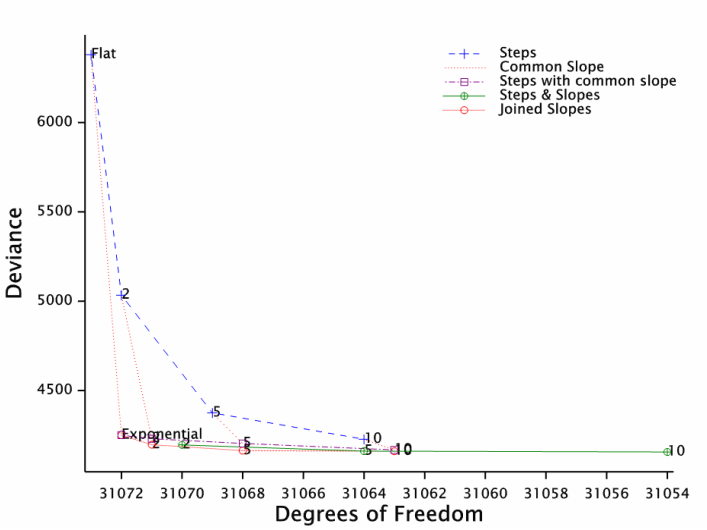
Steps and slopes can be combined, but this loses the practical advantages of slopes forming a continuous function, and is profligate in degrees of freedom. It is included for analysis to compare the influence of steps and slopes.

No attempt has been made in this analysis to find a parsimonious model by combining bands.

4.3.2 Fit

The statistical fit of alternative empirical functions is shown in figure 4.11 and with greater detail in figure 4.12.

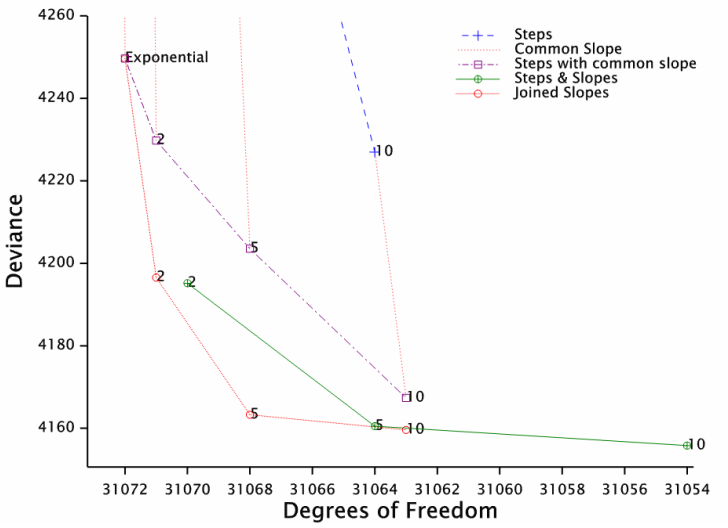
Figure 4.11 Fit of empirical functions



Steps alone give a relatively poor fit. The simplest case of just two levels is very crude, in effect classifying all trips into short or long. This still captures over 60% of the fit of the Exponential model with a single slope. Only the 10-step model fits better than this, by a small margin, and a single slope always improves the step models greatly.

Although steps with a common slope improve on the single-slope Exponential significantly, sets of joined slopes fit much better still. Improvement from 5-slope to 10-slope is not significant ($3.68 \sim \chi^2_5$). Similarly, the improvement of adding steps between slopes is not significant ($2.77 \sim \chi^2_4$ for five bands), nor is the improvement from 5 to 10 bands with both slopes and steps ($4.71 \sim \chi^2_{10}$).

Figure 4.12 Fit of empirical functions – detail



4.3.3 Fitted models

4.3.3.1 Steps

There is a strong underlying downward trend in the deterrence functions. To show the distinctions within this overall trend, steps and slopes are expressed as the changes between adjacent bands at the breakpoints in most of the following tables. This parameterisation is achieved by step functions which remain at 1 beyond their upper bound (factors return to 0), and slope functions that continue to rise beyond the upper bound, rather than level off (see figure 4.10).

Even so, table 4.7 shows that each step is downwards and significant. Figure 4.13 suggests that much of the improvement in fit from extra divisions is simply from a better approximation to a continuous function.

Figure 4.13 Deterrence functions – steps

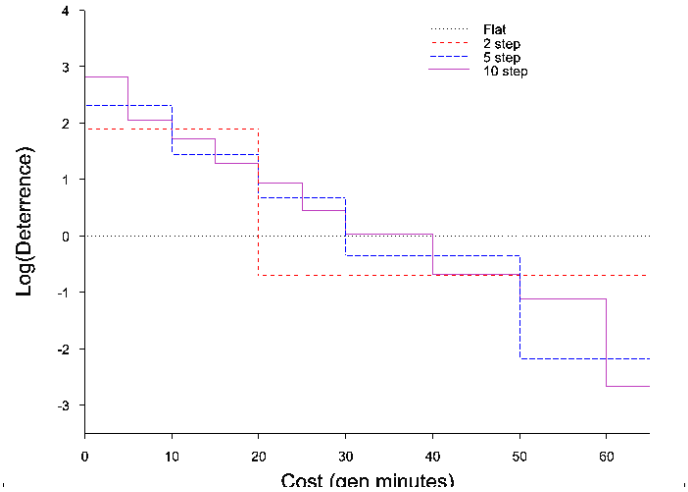
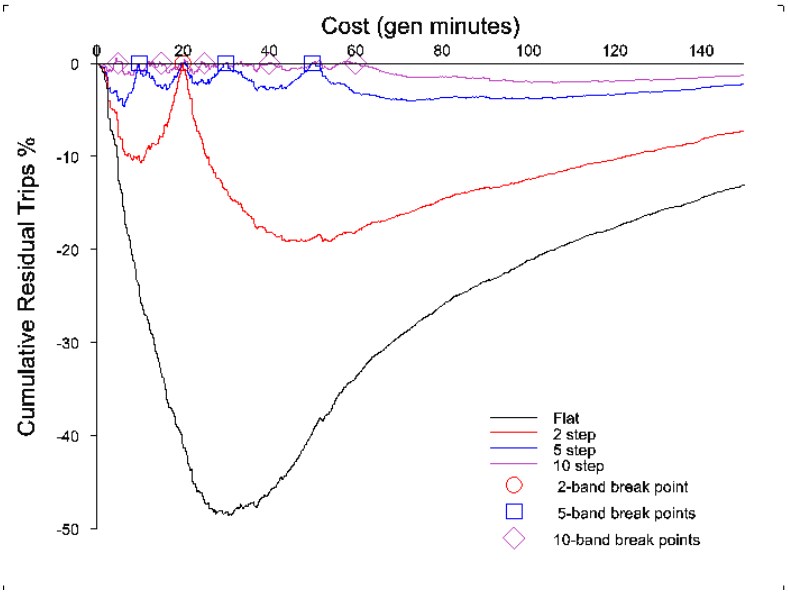


Table 4.7 Fitted coefficients – steps

Break points (g min)	2 band			5 band			10 band		
	coefficient	standard error	t statistic	coefficient	standard error	t statistic	coefficient	standard error	t statistic
5							0.76	0.12	6.2
10				0.88	0.09	9.5	0.34	0.12	2.8
15							0.43	0.12	3.5
20	2.60	0.071	36.8	0.77	0.10	7.7	0.34	0.13	2.6
25			*{36.7}*				0.49	0.15	3.3
30				1.03	0.11	9.4	0.42	0.15	2.8
40							0.71	0.16	4.6
50				1.83	0.13	14.4	0.44	0.18	2.4
60							1.55	0.20	7.9

{Italics in brackets are the square root of deviance changes corresponding to the t statistic above}

Figure 4.14 Cumulative residuals – steps



NB Wider vertical range than cumulative residual plots in figures 4.16, 4.18 and 4.20.

Because there is a constant (K) factor fitted for each band, the model reproduces the observed trips in each band. This causes the cumulative residual plots to touch the axis at each break point in figure 4.14. However, since no cost coefficient (L factor) is fitted for a slope, the total observed costs are not reproduced, so the plots do not have to be balanced either side of the axis. The plots lie below the axis, indicating underestimation at lower costs and overestimation at higher costs within each band. This demonstrates the need for a decreasing function within each band and an inadequacy of the flat-topped step model.

Table 4.8 Total costs in step models

Model	Fitted	Difference	
Observed	4,451,140	~	%
Flat	12,707,921	8,256,781	+185
2 step	8,266,558	3,815,418	+86
5 step	5,363,279	912,139	+20
10 step	4,847,912	396,772	+9
Exponential	4,451,161	21	+0.00047

Table 4.8 shows that as cost bands become narrower, the reproduction of total costs improves. In each band, the range of possible costs decreases, while the number of observed trips is always replicated by the K factors. The flat model can be seen as the extreme case of a step model, while the Exponential demonstrates the close fit that can be achieved.

Table 4.9 Total costs in 10-step model – by band

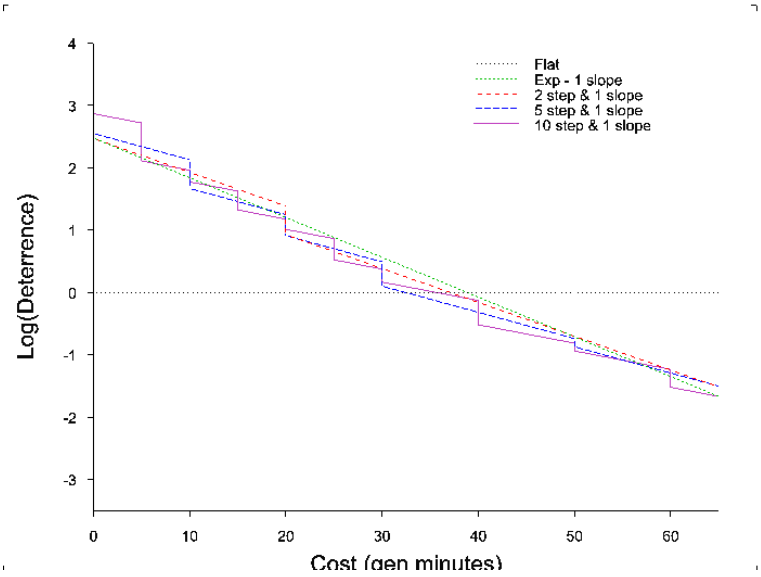
Band	Observed	Modelled	Difference	%
0–5	84,146	87,018	2872	3.41
5–10	196,802	203,133	6331	3.22
10–15	299,307	302,108	2801	.94
15–20	393,180	397,450	4270	1.09
20–25	491,230	495,256	4026	.82
25–30	352,335	354,665	2330	.66
30–40	653,053	657,312	4259	.65
40–50	530,367	538,389	8022	1.51
50–60	520,053	522,416	2363	.45
60 +	930,668	1,290,163	359,495	38.63
Overall	**4,451,140**	**4,847,912**	**396,772**	**8.91**

Table 4.9 shows that costs are overestimated in every band of the 10-step model, corresponding to the nett areas below the axis of the cumulative plot. However, most of the overall error occurs in the highest band, over 60 generalised minutes. This has the widest range, up to 276 minutes, while other bands are only 5 or 10 minutes wide.

In order to replicate costs well in a classic empirical model like this, such wide bands need to be subdivided, even if they contain few observed trips (6% in this case). This limits the scope for error in cost. In terms of the cumulative residuals in figure 4.14, breakpoints pin the curve to the axis, minimising the nett area between them. The high significance of the last steps in table 4.7 also supports this approach.

4.3.3.2 Steps with common slope

Figure 4.15 Deterrence functions – steps with common slope



The deterrence functions with steps in figure 4.15 show a general concavity, lying below the single slope Exponential in the middle and above it at the ends.

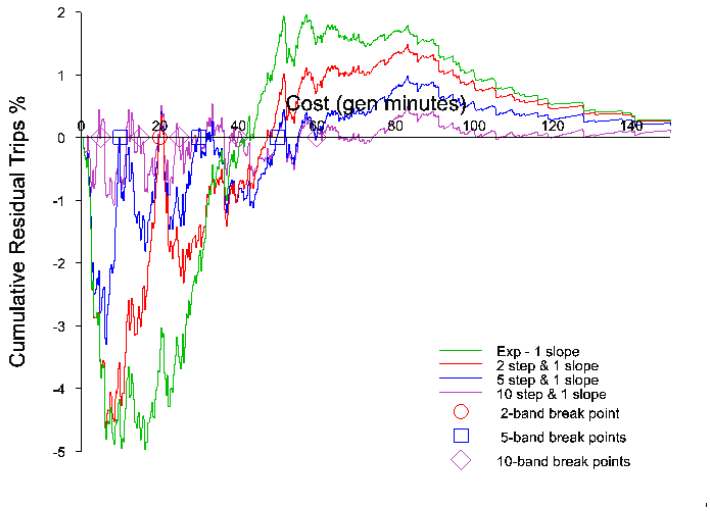
However, all the steps are downwards and table 4.10 shows that many of the coefficients are significant. The fitted cost coefficients for slopes, shown at the bottom of the table, are highly significant. They become less steep with more bands; the single slope Exponential has a coefficient of 0.0638. This suggests that the slope is being determined by the wide range of costs in the uppermost band – at least 60 to 160 generalised minutes – and the steps then adjust to fit the other narrower bands. Again, it might be better to subdivide this upper band further.

Table 4.10 Fitted coefficients – steps with common slope

Break points (g min)	2 band			5 band			10 band		
	coefficient	standard error	t statistic	coefficient	standard error	t statistic	coefficient	standard error	t statistic
5							0.61	0.13	4.8
10				0.46	0.10	4.6	0.19	0.12	1.5
15							0.30	0.13	2.4
20	0.46	0.10	4.5	0.34	0.11	3.1	0.16	0.13	1.3
25			*{4.5}*				0.34	0.15	2.3
30				0.39	0.12	3.2	0.21	0.15	1.4
40							0.39	0.16	2.4
50				0.13	0.17	0.8	0.12	0.19	0.7
60							0.29	0.24	1.2
Slope (gen min^{-1})	0.0543	0.0029	18.5 *{28.4}*	0.0421	0.0045	9.4 *{13.1}*	0.0296	0.0049	6.1 *{7.7}*

Italics in brackets are the square root of deviance changes corresponding to the t statistic above

Figure 4.16 Cumulative residuals – steps with common slope



Once a single cost coefficient (L factor) is added to fit a common slope, the total costs are fitted over the whole model, but not necessarily in individual bands. The cumulative residual plots balance about the whole axis, and are still forced to touch it at each breakpoint by the K factors reproducing trips in each band. There is still underestimation at low costs and overestimation at high.

4.3.3.3 Joined slopes

Figure 4.17 Deterrence functions – joined slopes

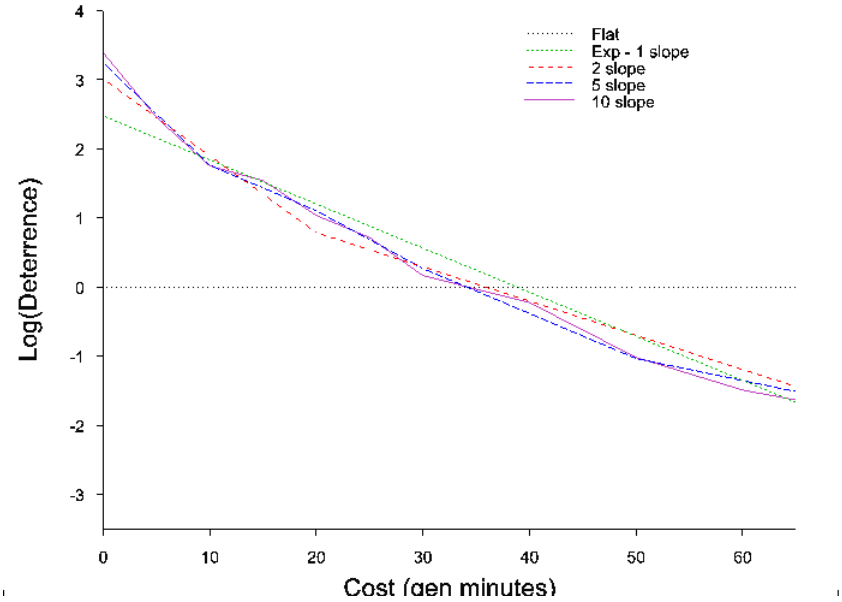


Table 4.11 Fitted coefficients – joined slopes

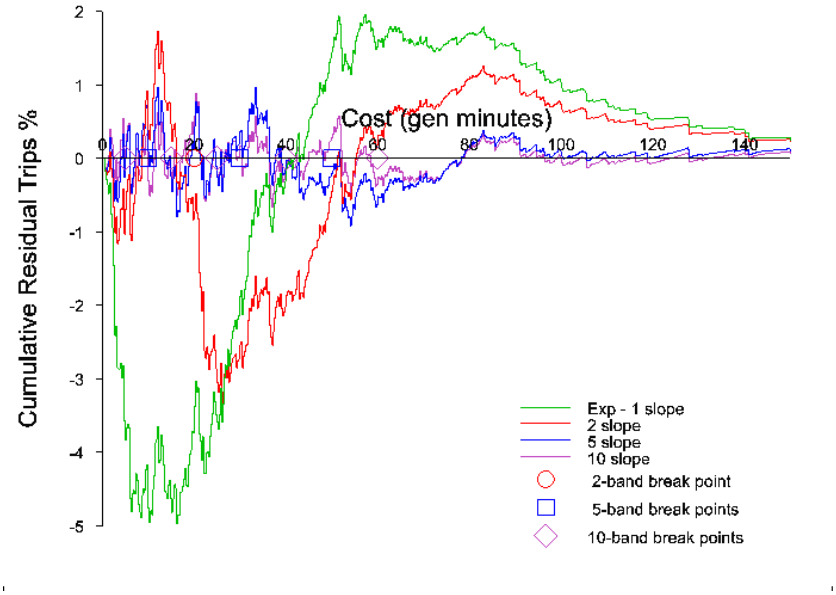
Break points (g min)	2 band			5 band			10 band		
	coefficient	standard error	t statistic	coefficient	standard error	t statistic	coefficient	standard error	t statistic
0 – 5	**0.1111**	**0.0068**	**16.3**	**0.148**	**0.020**	**7.3**	**0.188**	**0.062**	**3.0**
5							–0.048	0.088	–0.5
10				–0.082	0.031	–2.7	–0.098	0.066	–1.5
15							0.058	0.065	0.9
20	–0.0616	0.0083	–7.4	0.018	0.025	0.7	–0.037	0.069	–0.5
25			*{7.3}*				0.046	0.073	0.6
30				–0.019	0.021	–0.9	–0.071	0.058	–1.2
40							0.040	0.043	0.9
50				–0.033	0.010	–3.2	–0.032	0.045	–0.7
60							–0.018	0.027	–0.7

Bold figures are absolute values

Other figures are differences between adjacent bands at the breakpoint

Italics in brackets are the square root of deviance changes corresponding to the t statistic above

Figure 4.18 Cumulative residuals – joined slopes



Although the overall shape of curves in figure 4.17 is concave, they are not consistently so. However, table 4.11 shows that none of the increases in slope, with a positive change in coefficient that goes against the concave pattern, are significant.

Because there are no constant K factors for each band, the models do not reproduce the number of trips in each band, and the curves in figure 4.18 do not touch the axis at the breakpoints. Although there are separate slopes (L factors) for each band, observed costs are not reproduced for each band, presumably because the individual slopes are constrained to meet adjacent slopes at their breakpoints. Overall costs and trips are reproduced, since the model incorporates the basic Exponential form.

4.3.3.4 Steps and slopes

Figure 4.19 Deterrence functions – steps and slopes

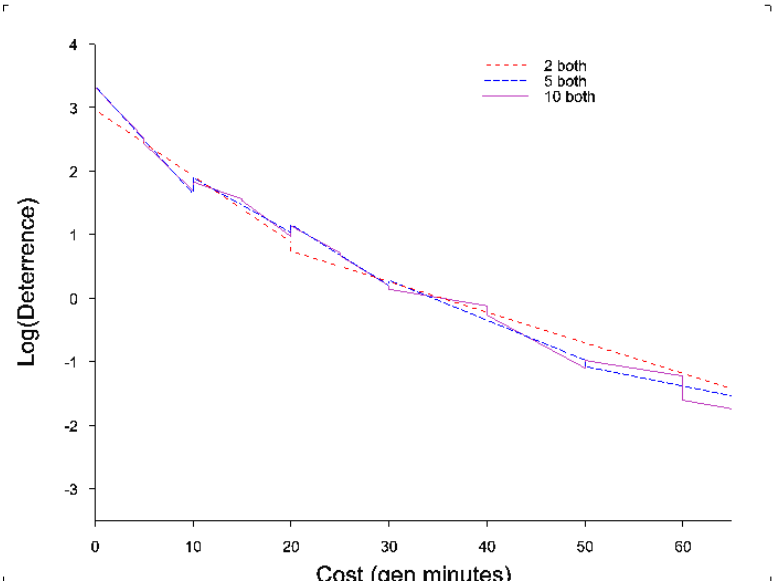


Table 4.12 Fitted coefficients – steps and slopes

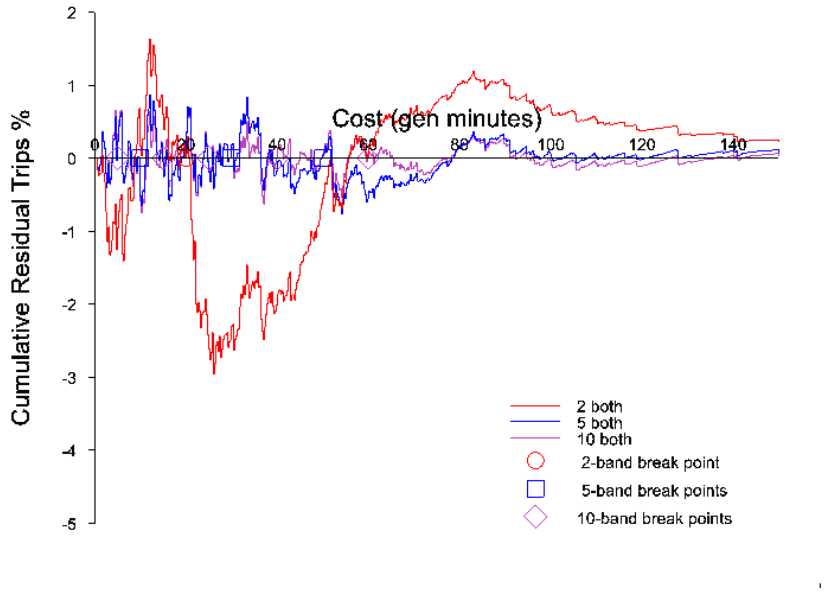
Break points (g min)	2 band			5 band			10 band		
	coefficient	standard error	t statistic	coefficient	standard error	t statistic	coefficient	standard error	t statistic
Steps – differences between bands									
5							0.09	0.25	0.4
10				−0.25	0.17	−1.5	−0.14	0.24	−0.6
15							0.03	0.24	0.1
20	0.14	0.12	1.2	−0.13	0.18	−0.7	−0.17	0.24	−0.7
25			*{1.2}*				0.04	0.28	0.2
30				−0.08	0.21	−0.4	0.05	0.30	0.2
40							0.15	0.30	0.5
50				0.10	0.21	0.5	−0.12	0.35	−0.3
60							0.38	0.34	1.1
Slopes – differences between bands									
5							−0.011	0.099	−0.1
10				−0.083	0.034	−2.4	−0.095	0.084	−1.1
15							0.059	0.083	0.7
20	−0.056	0.010	−5.9	0.009	0.033	0.3	−0.028	0.087	−0.3
25			*{5.9}*				0.010	0.101	0.1
30				−0.033	0.029	−1.2	−0.069	0.087	−0.8
40							0.057	0.053	1.1
50				−0.032	0.014	−2.3	−0.059	0.063	−0.9
60							0.002	0.047	0.0
Slopes – absolute									
0–5	**0.104**	**0.009**	**11.5**	**0.169**	**0.026**	**6.6**	**0.160**	**0.081**	**2.0**
5–10							**0.149**	**0.062**	**2.4**
1–15				**0.087**	**0.021**	**4.1**	**0.054**	**0.056**	**1.0**
15–20							**0.113**	**0.061**	**1.8**
20–25	**0.048**	**0.003**	**16.6**	**0.096**	**0.026**	**3.8**	**0.085**	**0.060**	**1.4**
25–30							**0.095**	**0.080**	**1.2**
30–40				**0.063**	**0.013**	**4.9**	**0.026**	**0.032**	**0.8**
40–50							**0.083**	**0.041**	**2.0**
50–60				**0.031**	**0.004**	**7.6**	**0.025**	**0.047**	**0.5**
60–							**0.027**	**0.005**	**5.6**

Bold figures are absolute values

Other figures are differences between adjacent bands at the breakpoint

{Italics in brackets} are change of deviance statistics corresponding to the t statistic above

Figure 4.20 Cumulative residuals – steps and slopes



The addition of K constants causes the cumulative plots to touch at breakpoints, while the L coefficients force the plots to balance about the axes between each pair of breakpoints.

Figure 4.19 again shows general concave forms, but with a mixture of upward and downward steps. The top part of table 4.12 shows that none of the steps are significant, and the middle part shows that the only differences in slope that are significant follow the concave pattern, with reduced slope at increased costs.

In models with joined slopes, the slopes are constrained to form a continuous function, so, by attachment to adjacent line segments, they reflect the overall downward trend as well as any trend within their band. With the introduction of steps freeing the ends of slopes, slopes are determined only by deterrence effects within their bands. The bottom part of table 4.12 shows that deterrence effects are significant within each band of the two- and five-band models. This confirms the deterrence effects within bands suggested by the cumulative residuals from step models in figure 4.14.

The slopes within the bands of the 10-band model are not clearly significant, probably because they are too narrow, except for the wider final band. This again suggests merit in higher breakpoints for a classic empirical model based on steps.

4.3.4 Statistics of fit – change in deviance and the t statistic

While change in deviance can test the introduction of several variables and degrees of freedom, the t statistic can only apply to one variable at a time. Where the number of bands is increased by more than one, the two statistics are not readily comparable.

Single variable changes occur when common slopes are added to step functions, shown in the bottom line of table 4.10. Here the change of deviance statistic is higher than the t statistic.

Single variable changes also occur at the breakpoint in two-band models. The change of deviance statistics shown for these in tables 4.7, 4.10, 4.11 and 4.12 compare closely with the t statistics, even though some statistics are higher than for the common slopes.

4.4 Geographic segmentation

The WTSM internal study area is divided progressively into:

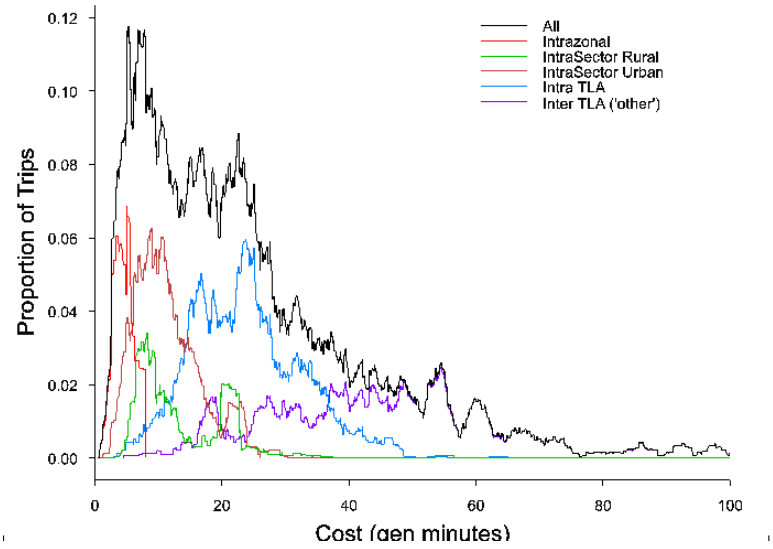
- seven territorial local authorities (TLAs)
- 15 sectors, classified as urban or rural
- 225 zones.

This leads to a geographic segmentation of movements:

- inter TLA between different TLAs
- intra TLA within one TLA, but between different sectors
- intrasector within one sector, but between different zones
- intrazonal within one zone.

This segmentation is described in detail in TN17.1, section 2.4. The distribution of trip costs by segment is shown in figure 4.21.

Figure 4.21 Cost distribution by geographic segment



This shows that each segment has a distinct range of typical costs, but there is also considerable overlap. Mean trip costs by segment appear in tables 4.13 and 4.14.

4.4.1 K and L factors

K factors are separate constants applied to different parts of a matrix, such as trips that cross a river. The application of a K factor causes the fitted model to replicate total observed trips over its scope. If the scope of a K factor is just a set of productions or attractions (complete rows or columns in the matrix), its effect in simple distribution will be absorbed by the balancing factors.

L factors are separate cost coefficients, or parameters in WTSM terminology, applied to subsets of the matrix. The application of an L factor causes the fitted model to replicate total observed trip costs over its scope. K and L factors with common scopes will thus replicate mean trip costs within the scopes.

The WTSM takes geographic segments as scopes for both K and L factors in HBW distribution.

K and L factors are the equivalent to the steps and slopes of empirical models in the previous section. In empirical models their scopes are the cost bands. In this single dimension of cost, separate slopes (L factors) can be parameterised to meet at the boundaries and form a continuous deterrence function. In more complex segmentation, it is hard to avoid steps at the boundaries between scopes, even if L coefficients are applied as well as K constants.

4.4.2 WTSM formulation

In the WTSM, HBW distribution is undertaken jointly with mode split and much of the formulation is designed to accommodate the mode split.

Productions are segmented by household on car availability. The principal household segments making car trips are 'competition' with more drivers than cars, and 'choice' with a car for every driver. These share the generalised costs described at the start of this chapter. The few car trips from 'captive' households, without cars, are also included in these analyses. For this segment, the WTSM uses a different cost structure, based on public transport costs.

Slow modes (walk and pedal-cycle) are included in the car distribution for the competition segment.

External zones are treated as separate TLAs, so trips to and from them are appear as inter-TLA in the WTSM. They are excluded from these analyses.

The K factors fitted to intrazonals are additional to the intrasector factors, so the total factors shown in table 4.13 differ between urban and rural for the competition segment. A common K factor is fitted to urban and rural intrazonals in the choice segment.

K factors have an arbitrary base level. This is inter-TLA ('other') in TN17.1 tables, and is adopted in these analyses. There is also a modal constant, which is included in all K factors in table 4.13, appearing as the inter-TLA K factor. It differs between competition and choice, but the trips in the segments are constrained by separate sets of production trip ends. However, these sets are common to car and public transport modes, and the modal constant affects the split between them.

Intrazonals have no separate L factors; by their nature there should be a limited range of costs for intrazonals, which are capped at five generalised minutes in the WTSM. The L factors for intrasectors (urban or rural) apply.

Table 4.13 WTSM K and L factors

Geographic segment	Mean cost (gen min)	Household segment			
		Competition		Choice	
		K, constant	L, parameter	K, constant	L, parameter
Intrazonal urban	2.31	4.141	0.1175	2.591	0.0834
Intrazonal rural	3.63	3.716	0.0899		0.0802
Intrasector urban	10.05	3.47	0.1175	2.346	0.0834
Intrasector rural	12.59	3.045	0.0899		0.0802
IntraTLA	22.46	2.982	0.0857	1.861	0.0587
InterTLA ('other')	48.80	1.6	0.0445	0.889	0.0391

Source: TN17.1, tables 3.2–3.4, files \311\HBWDMconst & ~gamma.311

Both K and L values tend to decrease down the table. Intrasector rural and intraTLA are similar in the competition segment. In the choice segment, urban and rural intrasector L factors are similar – they already share a common K factor.

The competition segment tends to have higher L factors, indicating a greater sensitivity to travel cost. This is also apparent in the greater slope of the functions in figure 4.22.

Figure 4.22 WTSM deterrence function – competition household segment

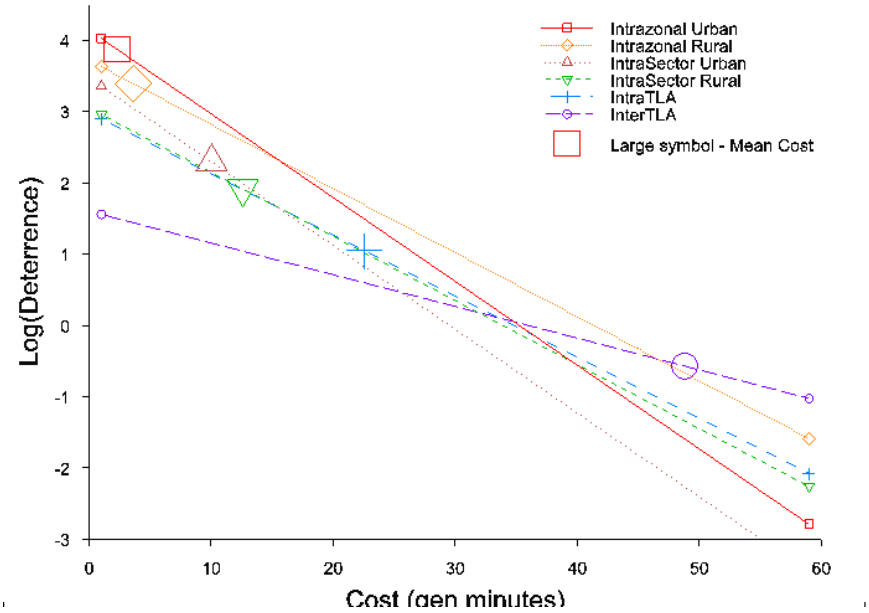
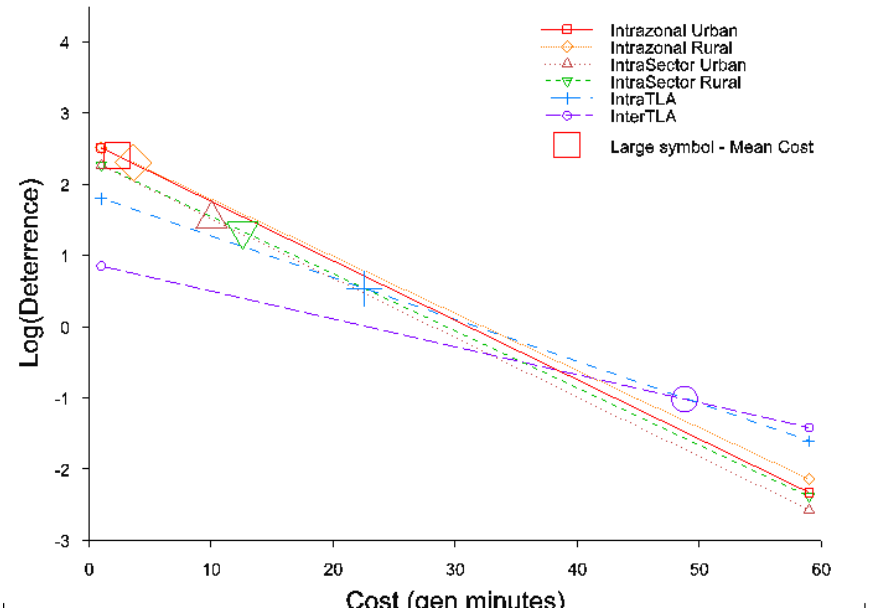


Figure 4.23 WTSM deterrence function – choice household segment



Figures 4.22 and 4.23 show the WTSM deterrence functions for the two main household segments. The mean cost for each segment is also plotted where it crosses the line. This gives a rough indication of the line's operating area, a summary of the distributions shown in figure 4.21. In both plots, the trace of these mean points suggests a reduced slope at higher costs.

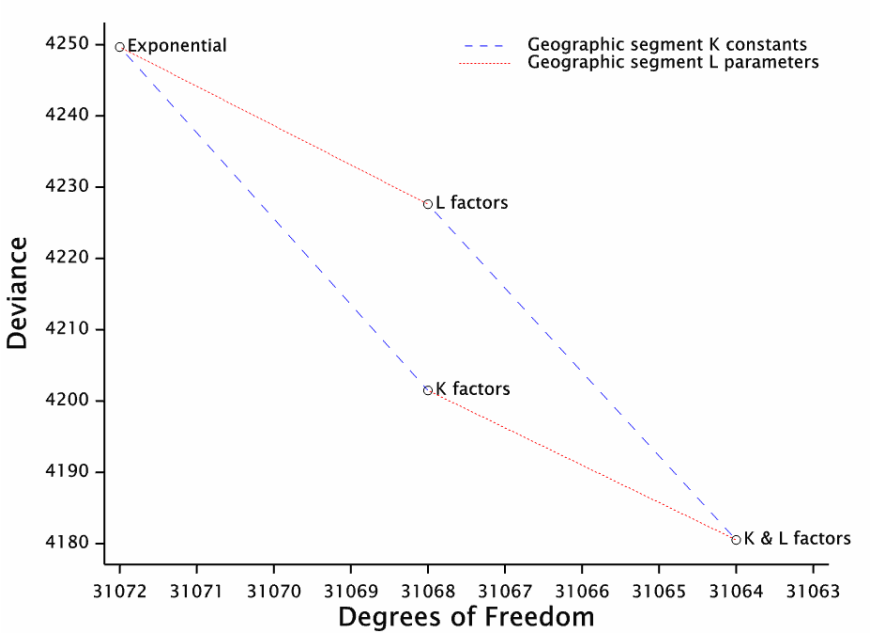
4.4.3 Fit

K and L factors are treated as complete sets, without trying to find a parsimonious model by combining geographic segments.

Intrazonals are fitted as a single separate segment, for both urban and rural zones, with both K and L factors. The small range of intrazonal costs will limit the fitting of their L factor.

The improvement in fit with the addition of these factors is shown in figure 4.24.

Figure 4.24 Fit of geographically segmented K and L factors



The K factors clearly give a greater improvement in fit than the L factors. Both are very significant, with changes in deviance of 47.1 and 21.0 ($\sim\chi^2_4$) for K and L factors entering the combined model. The initial model at the top left of the plot is an Exponential model with a single cost coefficient, equivalent to a single common slope on the deterrence plots.

An even simpler K factor model is possible. Without any cost coefficient, deterrence is determined by the geographic segment alone and is not sensitive to the range of costs within each segment, which can be considerable. Not surprisingly, the fit is poor with a deviance of 4649 (31,069df), well off the top of figure 4.24. This is markedly worse than the equivalent flat-stepped empirical model, segmented on cost, with a deviance of 4375 for five steps (again 31,069df). Even so, it still amounts to 81% of the reduction in deviance achieved by the simple Exponential model. The deterrence function from this simplistic K factor model is plotted in figure 4.25.

4.4.4 Fitted models

Figure 4.25 Deterrence function – K factors, no cost coefficient

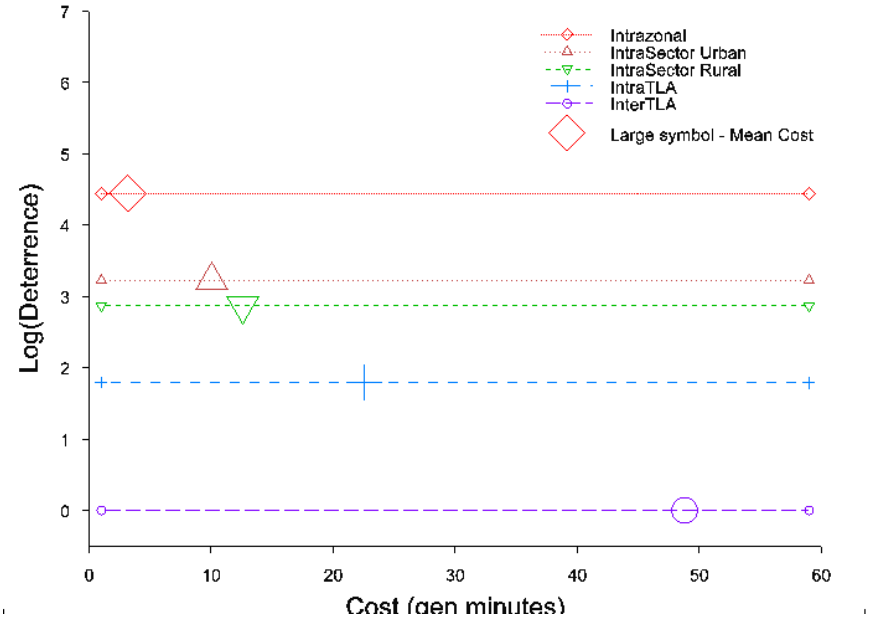


Figure 4.26 Deterrence function – K factors, single cost coefficient

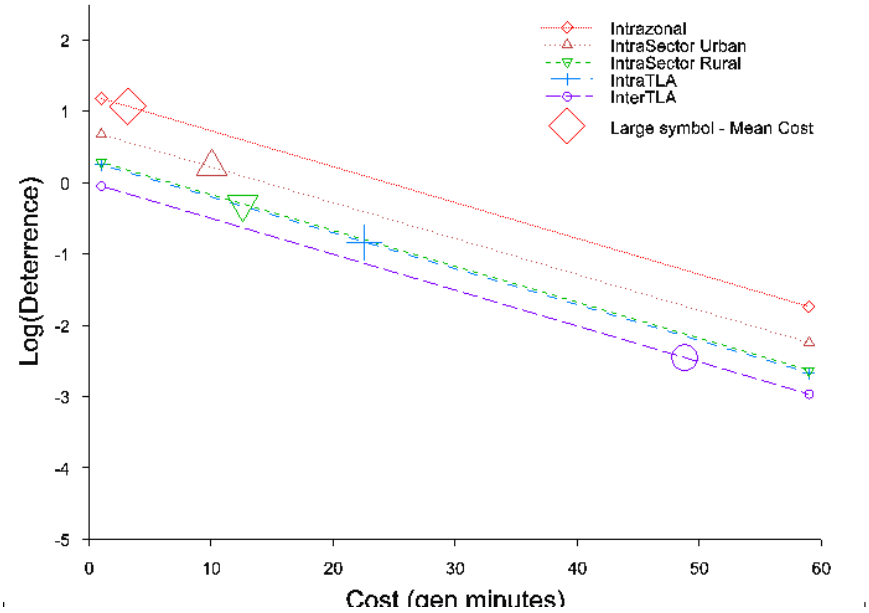
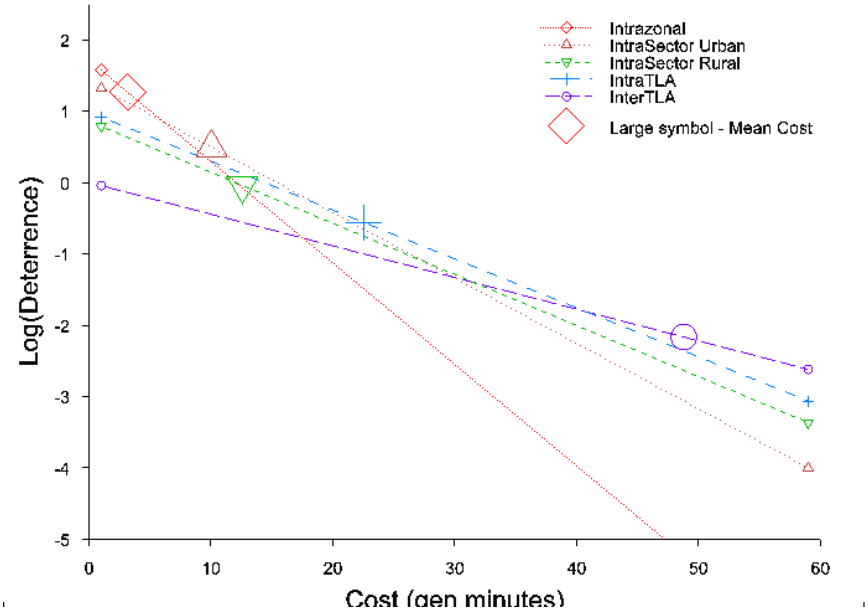


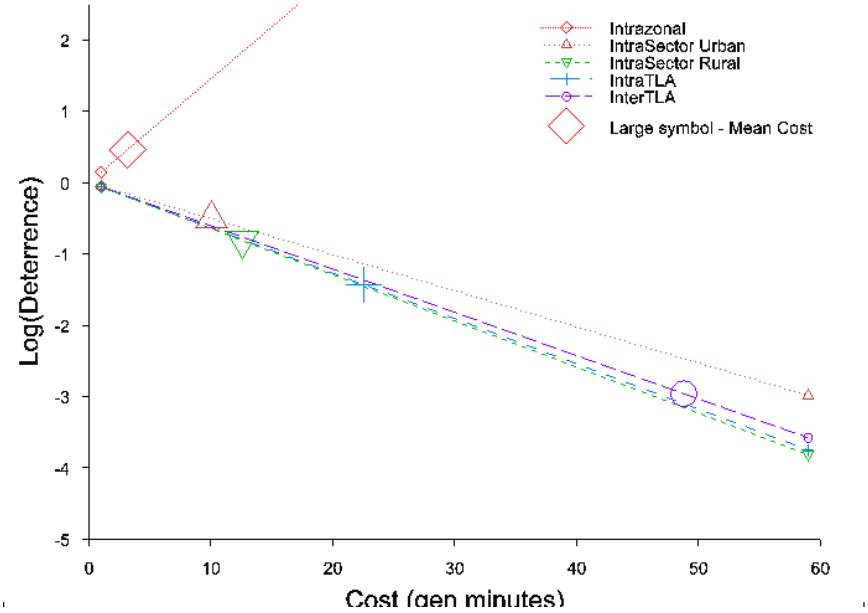
Figure 4.26 shows the deterrence function from K factors with a single cost coefficient, giving a common slope. The vertical origin is arbitrary, but the scale is the same as in other plots of deterrence functions.

Figure 4.27 Deterrence function - K and L factors



With L factors added in figure 4.27 each segment has a separate slope as well as intercept.

Figure 4.28 Deterrence function – L factors



When K factors are removed, there is a common intercept for all slopes. The counterintuitive but significant (2.6~t) slope for intrazonals only applies over their small range of costs; the errors induced must be compensated by better fit to other segments.

In all these deterrence functions, even K factors alone, the trace of mean costs not only forms a downward slope as would be expected, but also shows a diminishing slope for higher costs. The coefficients are shown in table 4.14.

Table 4.14 Fitted K and L factors

Geographic segment	Mean cost (gen min)	K only		K & λ		K & L		L	
		K	L	K	L	K	L	K	L
Intrazonal	3.16	4.443		1.228		1.716	0.142		−0.145
Intrasector urban	10.05	3.226		0.722		1.416	0.092		0.051
Intrasector rural	12.59	2.868	0	0.334	0.050	0.859	0.072	0	0.065
Intra TLA	22.56	1.800		0.300		0.986	0.069		0.064
Inter TLA ('other')	48.80	0.000		0.000		0.000	0.044		0.061

Figure 4.29 Cumulative residuals of geographic segmentation

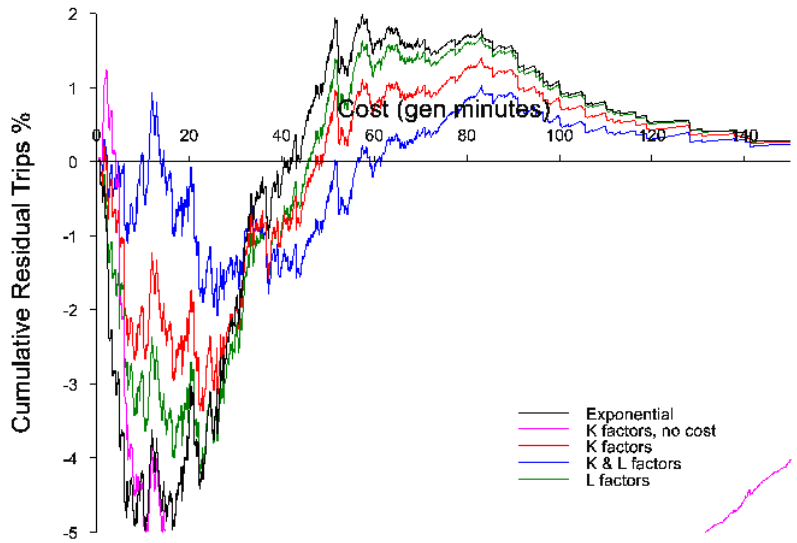


Figure 4.29 shows that adding K and L factors for five geographic segments does reduce the discrepancies of the Exponential model, but not as much as the five-slope model in figure 4.18. Plotting residuals against cost may favour empirical models, because they are segmented by cost, but it is not possible to plot geographic segmentation on a continuous axis. As in the deviance plots, the K constants are more useful than the L coefficients. However, without any cost coefficient, K constants alone fail to replicate the observed total trips costs, so this model's line is not balanced about the axis. The line falls out of the bottom of figure 4.29; it does eventually return to the axis via the bottom right corner of the figure because total trips are replicated by the constants.

With multiple factor levels and degrees of freedom, change in deviance and t statistics cannot be readily compared.

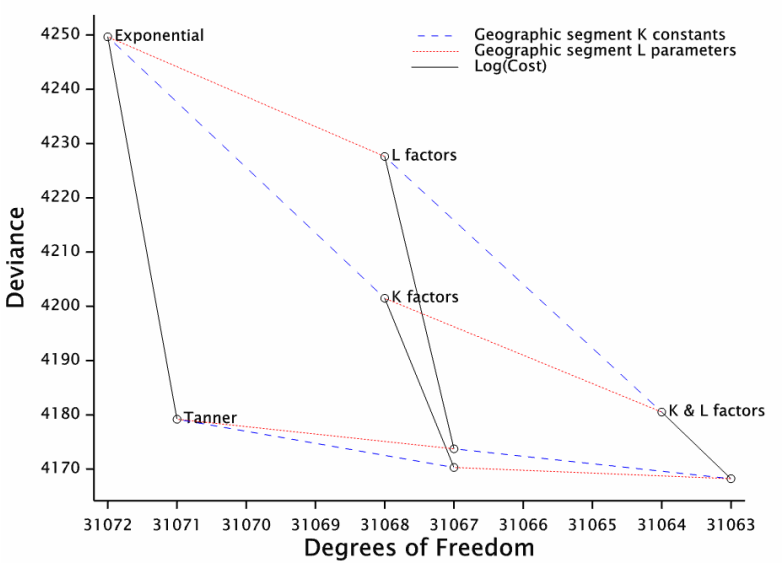
Some correlations between K and L factors are increased by taking the inter-TLA segment as the base level, since it has the highest typical costs furthest from the intercept. However, the alternative of taking intrazonals as a base level is unappealing to the practitioner.

4.4.5 Comparison with analytical forms

The fitted models above show a concave form similar to the Tanner. This suggests that the significance of geographic segmentation is due to the curvilinearity in the Tanner function, but it is still possible that there is a geographical effect beyond that function.

To test this, the logarithm of cost, which adds curvilinearity to the Exponential with a single term, is added to compare the relative effects on fit. These are shown in figure 4.30.

Figure 4.30 Comparison of fit – Tanner vs geographic



This shows that while log(cost) always improves markedly on geographic segmentation models, geographic segmentation adds little to the log(cost) in the Tanner function and is not significant.

Therefore in this dataset of commuter trips by car, most effects of geographic segmentation can be accounted for more generally as trip cost effects and represented more simply by the Tanner function.

4.5 Splines

In Genstat, GLMs can fit cubic splines whose curvature can change between line segments, similar to the change in slopes between cost segments in empirical functions.

Whereas empirical straight line segments meet with an instantaneous change in slope, cubic splines are continuous curves both in slope and acceleration (first and second differentials), so the changes across knot points (where the cubic component changes) are smooth. The cubic component of the piecewise polynomial is a step function with changes at the knot points. Generally a cubic smoothing spline has knots at all the distinct data points, but they can be specified at a reduced set of knot points for computational efficiency.

Splines minimise curvature in the line while fitting the data; the amount of curvature over the whole function is related approximately to the degrees of freedom. Technically a spline is the solution to minimising the weighted sum of the sums of squared residuals and a roughness penalty, the integral of the squared second differentials. The weighting given to the two components is controlled by the

specified degrees of freedom, and this sets the balance between smoothness (minimising roughness) and lack of fit (not being able to respond to sudden changes in the y values).

Whereas polynomials continue to curve beyond the limits of the data to which they are fitted, splines continue as straight lines since this minimises curvature.

Splines are constrained to be smooth and extra degrees of freedom help them progressively fit into corners in the data. Empirical functions, particularly the traditional step, can fit corners too well and spend degrees of freedom smoothing out the function.

Fitted splines are not readily described by parameters. In Genstat the splines have been abstracted as a set of points from the fitting process.

Higher order splines of cost were found to oscillate at high costs (see figure 4.36). This was thought to be due to fitting the relative sparse points in wide ranges. In an attempt to stabilise the functions, splines were also fitted to the logarithm of costs, which shrank the range of high costs. This gives two series:

Spline: splines of cost

order 0 is flat, order 1 is the Exponential function

Spline(Log): splines of log(cost)

order 0 is flat, order 1 is the Power function

Fitting the logarithm of cost gives the Power function. The Tanner formulation fits both cost and its logarithm, leading to two further series of splines. Each contains a first order component of the alternative (natural or logarithm) term in cost and they crossover with their first order splines as the Tanner function.

Tanner spline: log(cost) + splines of cost

order 0 is the Power function, order 1 is the Tanner function

Tanner spline(Log): cost + splines of log(cost)

order 0 is the Exponential function, order 1 is the Tanner function

Logarithms expand the range of lower costs and may make functions more sensitive to the treatment of intrazonals.

In Genstat, splines have to be fitted without absorbing groups, which makes their calculation times longer than other functions. Genstat also offers locally weighted regression, a form of moving average, but this failed to converge.

4.5.1 Fit

Splines can be computed up to a very high order. Computation was continued up to the order 50, in steps of 5 beyond 15.

Figure 4.31 Fit of splines

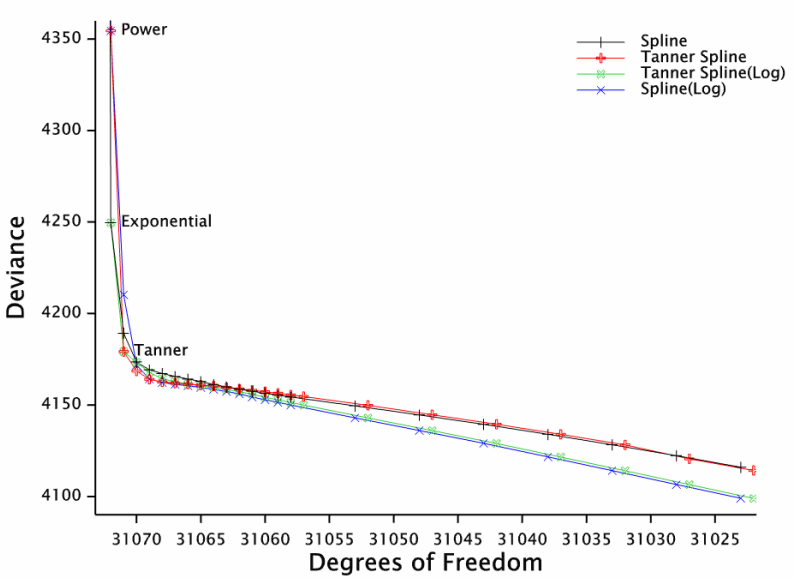
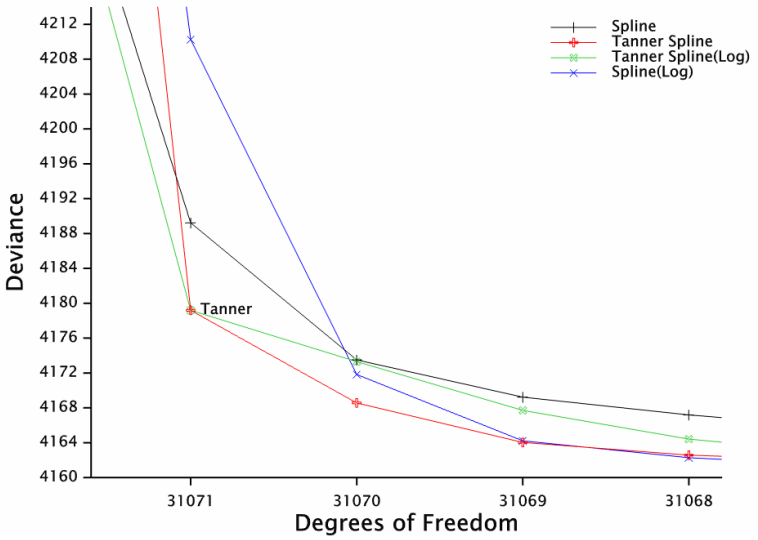


Figure 4.32 Fit of splines - detail



For two degrees of freedom (31,071 remaining), the Tanner function fits far better than second order splines of either cost or log(cost).

For a third degree of freedom (31,070 remaining), all four formulations improve on the fit of the Tanner. The best is the second order cost spline added to the Tanner, which is a very significant improvement ($10.6 \sim \chi^2_1$).

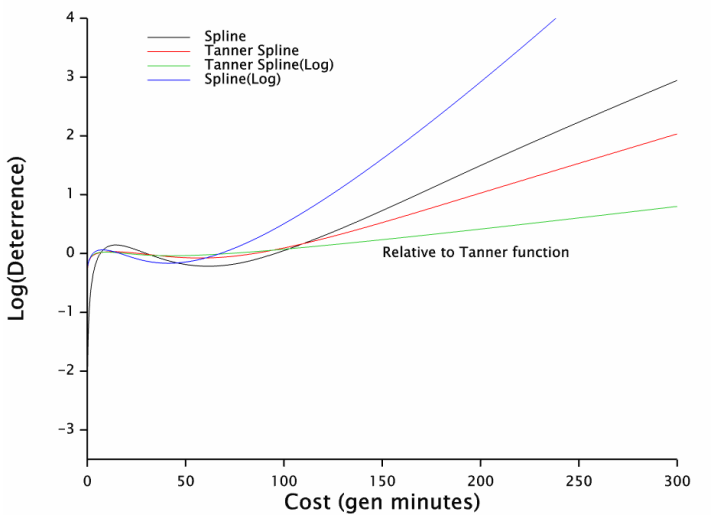
A third order cost spline added to the Tanner is still just the best fit for four degrees of freedom (31,069 remaining), and its improvement over the second order is significant ($4.5 \sim \chi^2_1$). Its residual deviance is almost matched by a fourth order spline of log(cost).

This formulation gives the best fit for higher orders, while Tanner plus a cost spline becomes the weakest. Splines of cost alone tend to give poorer fit, suggesting a benefit of including some element of log(cost). However, individual increases in the order are not significant.

4.5.2 Fitted models

No spline formulation fits as well as the Tanner function for two degrees of freedom. Other functions are plotted here with respect to the Tanner function, ie the amount by which their deterrence differs from the Tanner's. This helps emphasise the relatively small differences over the main cost range from 3 to 60 generalised minutes.

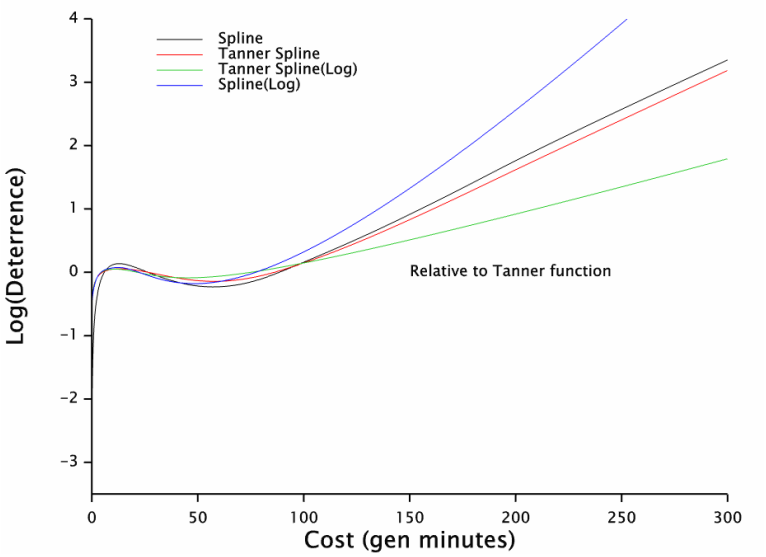
Figure 4.33 Splines with 3df



With three degrees of freedom, the main feature is an upward slope at higher costs, showing a reduced deterrence at these higher costs over and above the Tanner function, which allows curvature from the straight-line Exponential. Changes over the major operating range of the curve, between 3 and 60 generalised minutes, are relatively small, particularly the best-fitting Tanner splines of logarithms. Only the simple splines, which fit least well, depart markedly at very low costs.

There are relatively few trips, observed or fitted, in the upper range where splines depart markedly from the Tanner, so residual plots show little difference.

Figure 4.34 Splines with 4df



Adding a fourth degree of freedom, there is some convergence in the different formulations, but the overall pattern is the same.

By the 10th order oscillations appear in the upper end of splines of cost and in the lower end of the logarithmic formulations.

By the 50th order, differences between the Tanner and simpler formulations disappear except at the very ends of the functions. Both logarithmic and natural forms show oscillation in the middle range with similar forms. The principal effect is still an increase at costs above 80 generalised minutes and possibly a step in the function around that point.

Figure 4.35 Splines with 10df

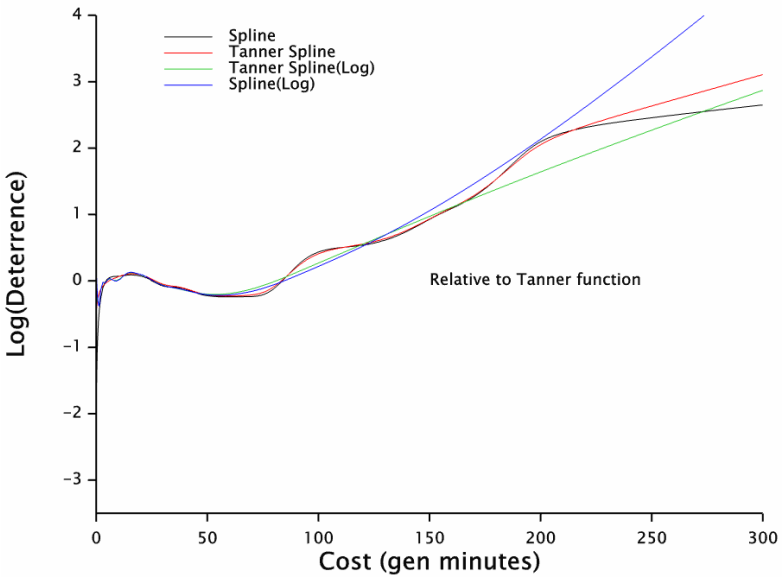
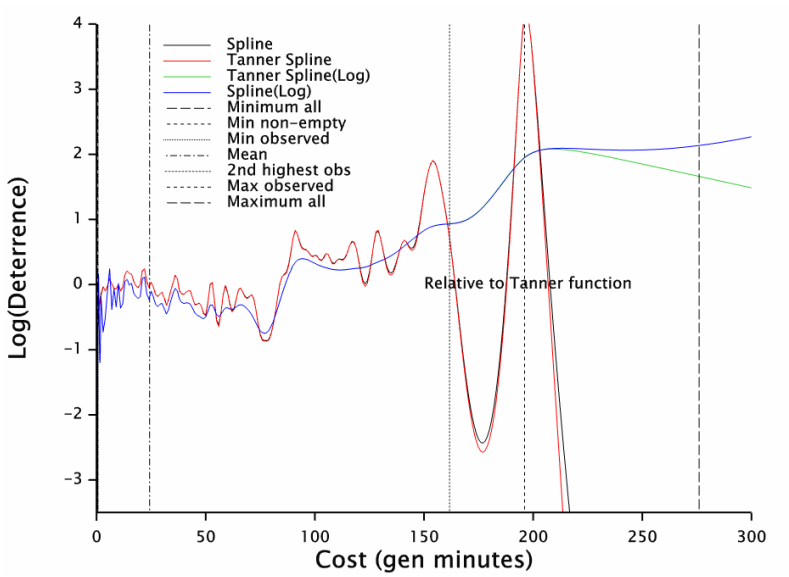


Figure 4.36 Splines with 50df



The final peak in the natural splines appears to be fitting the isolated highest observation, while deterring trips at higher costs where there are no observations, and in the interval from the second highest observation.

4.5.3 Turning points

As the order of the splines increases, turning points appear in the fitted functions. A first order turning point is where the curve flattens and then reverses the slope. This indicates increasing attraction with increasing travel cost, which is fundamentally implausible in economic terms and possibly unstable in iterative model fitting.

Turning points are not readily apparent in the figures above because the deterrences are plotted relative to the Tanner function. They are listed in the first part of table 4.15 as the lowest order of spline in which they occur and the generalised cost at which they occur – 'lower' below the mean cost of 25 generalised minutes, and 'upper' above.

Table 4.15 Turning points in splines

Base	Exponential		Tanner				Power	
Spline of	Cost				Log(cost)			
Range	Order	Cost	Order	Cost	Order	Cost	Order	Cost
First order: increase in travel with cost								
Lower	30	21	30	21	20	13	20	13
Upper	12	85	12	85	30	34	30	34
Second order in Exponential function: convex curvature in log(deterrence) vs cost								
Lower	14	23	14	23	9	15	9	15
Upper	8	90	8	90	14	37	14	37
Second order in Power function: concave curvature in log(deterrence) vs log(cost) – elasticity								
Lower	11	12	1	5	1	5	2	5
Upper	4	66	4	67	8	51	2	134

Second-order turning points can also be found. These depend upon the domain in which the curve is considered. In log(deterrence) vs cost, a straight line is the Exponential function and empirical evidence suggests a concave curve gives a better fit. Formulations where the curve becomes convex in this domain are shown in the second part of table 4.15.

In the domain of log(deterrence) vs log(cost), a straight line represents constant elasticity and the Power function. The better fitting Exponential and Tanner functions form convex curves in this domain, and formulations which produce concavity are given in the third part of table 4.15.

Once turning points appear in a spline series, they also occur for higher orders, or more turning points appear closer to the mean.

Turning points tend to appear first in natural splines in the upper ranges. The logarithmic form stabilises this, increasing the orders at which upper turning points appear, but at the expense of turning points appearing earlier in the lower ranges. The incorporation of a Tanner component has little effect in higher order splines.

The turning points appear well inside the practical range of costs.

While first-order turning points are clearly implausible, examination of the second-order turning points does not suggest any fundamental reason for rejecting the functions. Daly (2010) finds no such reason in his consideration of cost damping.

The first order turning points are for much higher-order splines than are significant, or are likely to be used in any pragmatic or parsimonious formulation.

4.5.4 Preferred form

A Tanner function with a second-order natural spline provides a parsimonious fit for one extra degree of freedom and is thus the most promising form for further analysis. It also leads to another well-fitting form in its third order.

A similar fit is provided by a fourth-order logarithmic spline. This is not a direct development of either the Exponential or Tanner functions and does not necessarily reproduce mean trip costs. However, it is the same form as a polynomial that fits well, though with two degrees of freedom rather than four. This form also fits well at higher orders of splines.

4.6 Polynomials

Fitting polynomials of the form

$$a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots + a_nx^n$$

is a more traditional approach to investigating curvature in regression lines. Unlike splines, their form can force them to continue curving beyond the limits of fitted data and makes them more likely to have turning points.

Odd-ordered polynomials are intrinsically more plausible, since they will ultimately tend to opposite directions at opposite ends of their ranges. However, some odd high-order polynomials fit in a counter-intuitive sense, low at low cost to high at high cost, with these extremities outside the main range of data.

No attempt was made to pick the best terms, ie particular powers in a polynomial series.

Although the addition of most terms reduced the residual deviance, some actually increased it slightly. This should not occur and indicates instability in the fitting process. This also appeared at higher orders in

divergences in parameters from nominally identical runs (with or without missing values appended to save the deterrence function) and finally in the failure in the fitting process at orders between 13 and 21.

The terms become highly correlated at these orders. Genstat does have a function (REG) for orthogonalising polynomial terms in simple cases. However, it only works for orders up to four, on the grounds that analysis of any higher orders requires very specialised knowledge and care. *Verb. sap.*

Four formulations involving log(cost) and Tanner bases have been fitted, the same as for splines.

4.6.1 Fit

Figure 4.37 Fit of polynomials

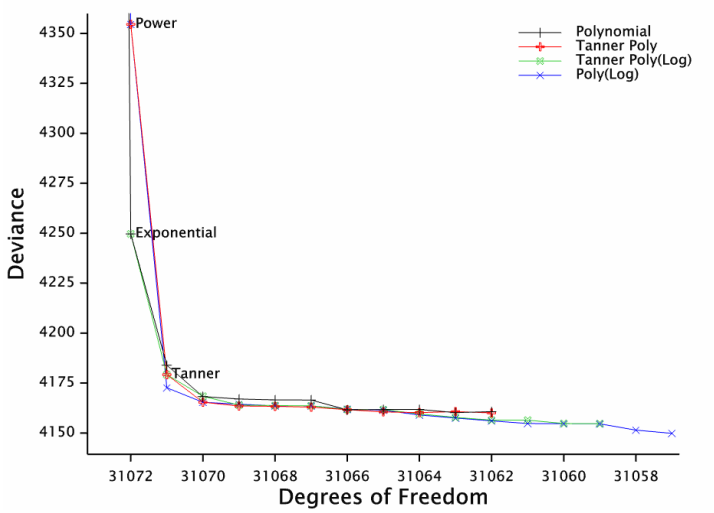
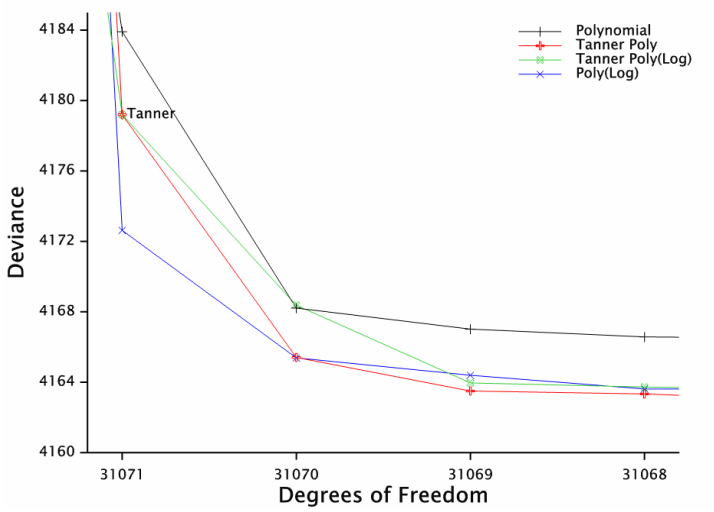


Figure 4.38 Fit of polynomials - detail



There is a slightly sharper 'knee' where the trace flattens out than for splines. Only improvements up to three degrees of freedom (31,070df remaining) are significant.

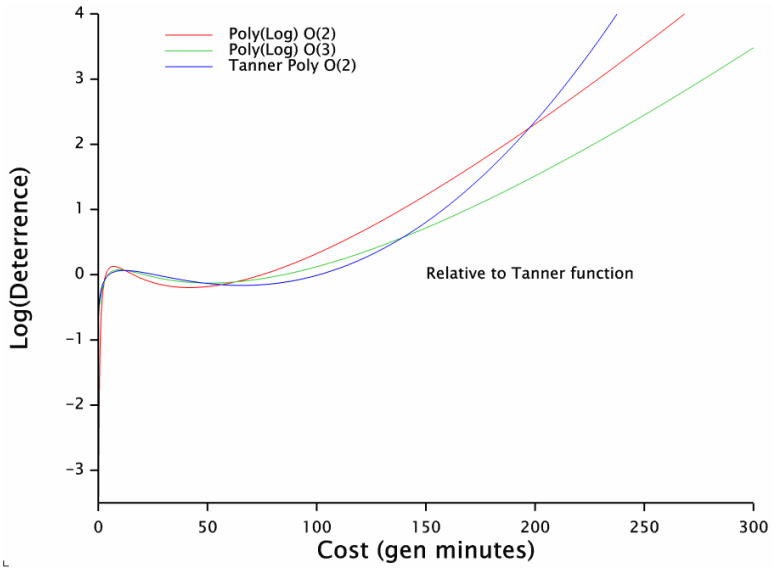
However, unlike splines, there is a function that fits better than the Tanner for the same two degrees of freedom (31,071df remaining). This is the polynomial of log(cost). The same form gives one of the best fits with three degrees of freedom, together with a Tanner polynomial of cost.

This form gives the best fit up to 7df, but by a narrow margin. At higher orders, logarithmic forms fit better.

4.6.2 Fitted models

The three functions fitting best for few degrees of freedom are shown in figure 4.39. They are plotted as differences from the Tanner function.

Figure 4.39 Best fitting polynomials



These again show the same form as the splines with increased trip probabilities above 100 generalised minutes, but relatively little difference below.

4.6.3 Turning points

First order turning points (changing from a decreasing function to an increasing one or vice versa) occur in low-order polynomials. These are implausible and potentially unstable in trip distribution synthesis. However, they are not surprising given the form of polynomials, since a second-order polynomial is a parabola. Turning points above and below the mean cost of 25 generalised minutes are shown in table 4.16.

Table 4.16 First order turning points in polynomials

Base	Exponential		Tanner				Power	
Polynomial of	Cost				Log(cost)			
Cost range	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Order	g min	g min	g min	g min	g min	g min	g min	g min
1	*	*	*	*	*	*	*	*
2	*	174	*	206	1	*	2	*
3	*	*	*	*	*	490	*	*
4	*	238	*	*	*	320	*	*
5	*	*	*	238	0.6	*	0.6	450

Base	Exponential		Tanner				Power	
Polynomial of	Cost				Log(cost)			
Cost range	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Order	g min	g min	g min	g min	g min	g min	g min	g min
6	*	262	*	*	*	*	*	360
7	*	167	*	168	*	*	*	*
8	*	166	*	169	1	226	*	*
9	*	166	*	166	3	*	1	228
10	*	170	*	169	1	226	3	*
11	*	167	*	167	1	242	1	228
12	*	169	*	167	2	*	1	*
13	~	~	*	167	2	256	2	*
14	~	~	~	~	2	220	2	*
15	~	~	~	~	~	~	2	169
16	~	~	~	~	~	~	1	163

Lower – below 25 generalised minutes and closest to it

Upper – above 25 generalised minutes and closest to it

* – no turning point in range 0.6 – 500 generalised minutes

~ – model did not converge

Shaded – best fitting models

Logarithmic forms reduce the occurrence of upper turning points, but at the expense of introducing lower ones.

Turning points tend to appear at the edge of the range of observed costs, unlike spline functions.

Even if turning points occur outside the current range of costs, increases in petrol costs or congestion could scale up generalised cost. If this pushed the range of costs beyond the range of calibration and above a turning point in the deterrence function, the trip distribution model could show a sudden increase in very long commuter trips.

4.6.4 Preferred form

Although more traditional for examining curvature in linear regression models, polynomials are less suitable than splines for investigating deterrence functions, where a monotonic decrease is expected.

Of the three well-fitting functions plotted in figure 4.39 the two second-order curves have turning points at the edge of their working ranges. The third-order curve is logarithmic and is not a direct development of either the Exponential or Tanner functions, but is included in later comparisons for contrast.

4.7 Other deterrence functions – non-linear fitting

The deterrence functions considered so far have all been linear within the log-linear GLM, or presented as linear functions by manipulation of parameters. For splines, the Genstat package internally calculated knot points akin to the break points between steps in empirical functions and then fitted polynomials between them.

Several rarer deterrence functions have been found (section 2.2.5 and Daly 2010), and some of these cannot be fitted directly by standard GLMs. However, Genstat offers extensions to optimise non-linear functions within GLMs, which allows some of these to be fitted. The general form of the deterrence function is exp[-λf(Cost,α,β,δ)], where λ and α,β,δ are linear and non-linear coefficients. f() is the non-linear function of cost and the non-linear coefficients, shown in table 4.17. The exponential is provided by the logarithmic link of the GLM.

As far as possible, the non-linear functions have been formulated for consistency amongst each other, using the symbols

- α for Power
- β for multiplication (or division)
- δ for addition or offset

to cost or a function thereof. However, in the presence of logarithms and exponents, these distinctions become blurred and the symbols have to do double duty in the double Power and Exponential functions. They represent the same coefficients as in the deterrence functions introduced in section 2.2.5, but the original sources identified there may use different symbols or express the function differently.

Table 4.17 Non-linear functions

Deterrence function	Source	Non-linear term, f(cost,α,β,δ)	df
Flat – minimal		*constant*	0
Exponential		*linear in* cost	1
Tanner		Cost + δ log(cost)	2
Power, or root log-normal ~ zero offset		*linear in* log(cost)	1
Root log-normal ~ unity offset	from below	*linear in* log(cost+1)	1
Root log-normal ~ fitted offset	from below	log(cost+δ)	2
Log-normal ~ zero offset	from below	*linear in* $(\log(\text{cost}))^2$	1
Log-normal ~ unity offset	OmniTrans	*linear in* $(\log(\text{cost}+1))^2$	1
Log-normal ~ fitted offset	from above	$(\log(\text{cost}+\delta))^2$	2
Top log-normal	OmniTrans	$(\log(\text{cost}/\beta))^2$	2
Box-Cox as Daly's Power formulation	Daly mechanism G	Cost^{α}	2
Box-Cox	VISUM	$(\text{Cost}^{\alpha}-1)/\alpha$	2
Box-Tukey	Daly mechanism Ea	$((\text{Cost}+\delta)^{\alpha}-1)/\alpha$	3
EVA1	VISUM	$\log(\text{cost}+1)/(\exp(\beta\text{cost})+\delta)$	3
EVA2 (or Schiller with λ fixed at unity)	VISUM	$\log((\text{cost}/\beta)^{\alpha}+1)$	3 (2*)
Double Power	Tmodel	$-\log(\text{cost}^{\alpha}+\delta\text{cost}^{\beta})$	3*
Double Exponential	Christchurch, TRACKS	$\log(\exp(-\alpha\text{cost})+\delta\exp(-\beta\text{cost}))$	3*

* linear coefficient λ fixed at unity; all fitted coefficients non-linear

Daly – mechanisms listed in table 1 of his 2010 review of cost damping

The flat model, with no cost effects, and the basic Power and Exponential models are included for comparison. The Tanner function can be expressed as a linear model with the two terms cost and log(cost), but it was also formulated as the non-linear function shown above.

The log-normal with a fixed offset of 1, as described in OmniTrans, is fitted as a linear model of the transformed cost. As the offset must be of the same dimension as cost, it must vary according to the units of cost, such as generalised minutes, dollars or cents. Therefore a non-linear model was formulated to fit the offset as the non-linear coefficient δ, and as a null case, a linear model with zero offset was also fitted.

The square of log(cost + offset) is formed in these log-normal models. As an intermediate, simpler stage, these models were also tested without the squaring, and termed 'root log-normal'. The case with zero offset is the Power deterrence function.

The offset distinguishing the Box-Tukey from the Box-Cox function is fitted as a second non-linear coefficient δ. The Box-Cox reduces to the Power or Exponential functions for α=0 or 1. Other relationships between the functions are set out in table 4.19 and indicated by lines joining the models in figures 4.40 and 4.41.

The final column of table 4.17 shows the degrees of freedom taken in fitting the deterrence function of cost, ie the number of fitted coefficients. This includes the linear coefficient λ (if not fixed), and excludes the 356 balancing factors for non-empty zones which are simultaneously fitted as linear coefficients.

4.7.1 Fit

The EVA and double functions at the bottom of table 4.17 did not converge. The deviances from the functions that were fitted are plotted in figures 4.40 and 4.41 and listed in table 4.18 with their fitted coefficients.

Figure 4.40 Fit of non-linear models

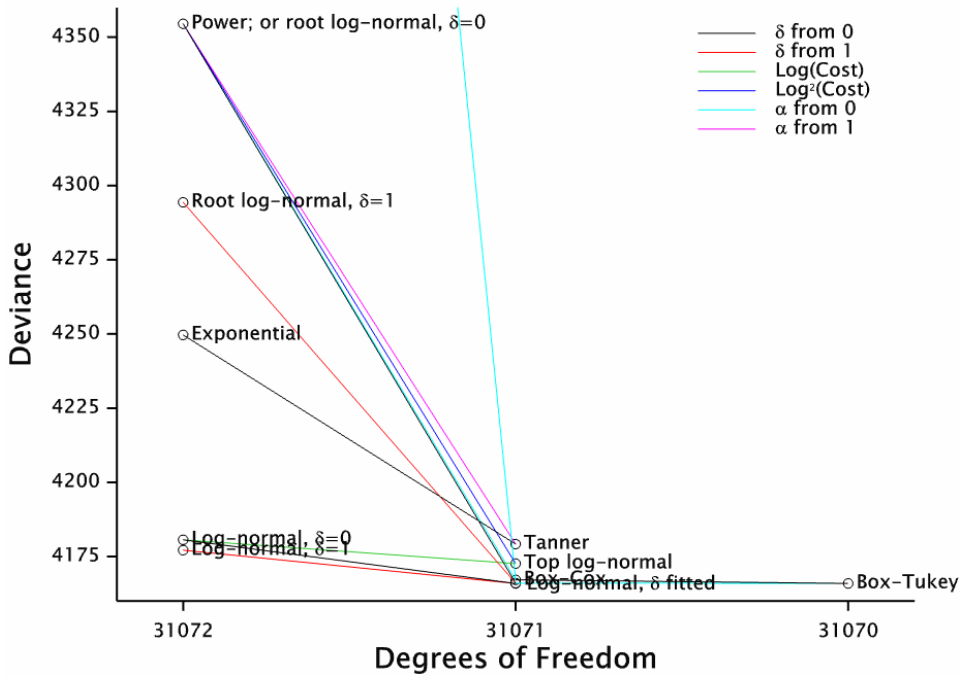
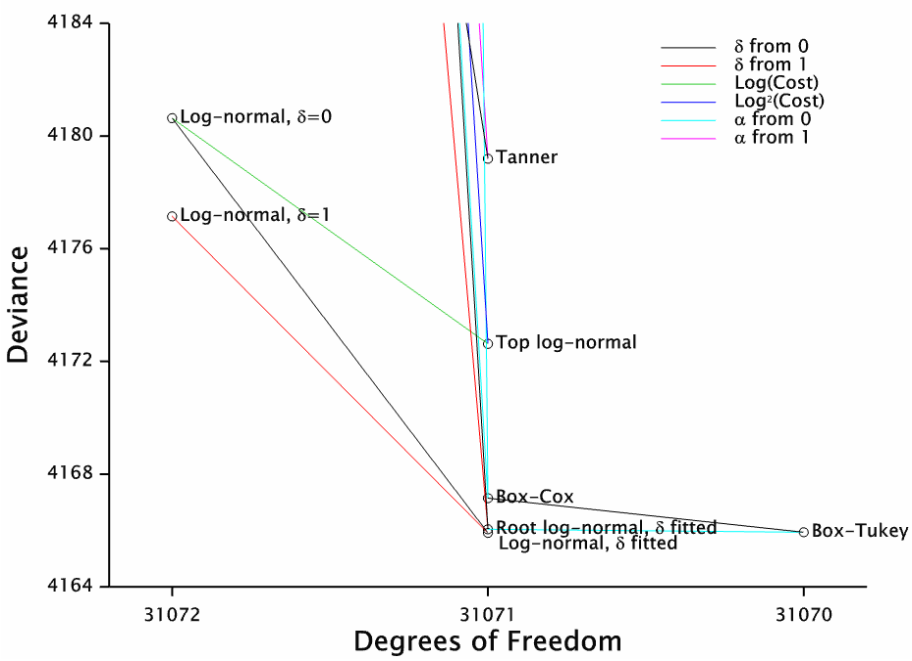


Figure 4.41 Fit of non-linear models – detail



Log-normal transforms with offsets δ fixed at either zero or unity can be fitted with a single linear coefficient. Their fit is not only markedly better than the single-coefficient Exponential model, but is similar to the two-coefficient Tanner model and better with an offset of unity. Calibrating the offset as a second non-linear coefficient of 11.6 generalised minutes gives a significant improvement, fitting better than any other model with the same degrees of freedom.

The root log-normal model fits almost as well with a larger fitted offset of 27 generalised minutes. With the offset fixed at unity, the fit is worse than the Exponential, though still better than the Power, which is the same model with zero offset.

The root log-normal with fitted offset can be seen as a special case of the Box-Tukey model with the Power α fixed to zero. The Box-Tukey is not a significant improvement on this root log-normal, nor on another special case, the Box-Cox, where the offset δ is set to zero. The fit of the Box-Cox is only a little worse than the log-normal or its root with fitted offset, but a significant improvement on the Exponential or Power, which are special cases for α set to unity or zero respectively. The non-linear Box-Cox also fits markedly better than the Tanner, which is a linear combination of the Exponential and Power.

The residual deviance of the top log-normal falls midway between those of the Tanner and the Box-Cox. The top log-normal's linear components of log(cost) and its square (the log-normal with zero offset) both give significant improvements when added to the other.

Table 4.18 Fitted non-linear models

Deterrence function	Residual deviance	Linear		Non-linear			Correlation with λ
		λ	se	coef.	value	se	
Flat – minimal	6380.44	0		~			
Exponential	4249.7	0.0638	0.0022	~			
Tanner *as non-linear*	4179.19	0.0364	0.0040	δ	18.191	4.159	-0.95
Root log-normal ~ zero offset	4354.50	1.416	0.033	*δ*	*0*	*fixed*	
Root log-normal ~ unity offset	4294.36	1.588	0.038	*δ*	*1*	*fixed*	
Root log-normal ~ fitted offset	4166.04	3.84	0.56	δ	27.03	7.35	0.97
Log-normal ~ zero offset	4180.63	0.2723	0.0073	*δ*	*0*	*fixed*	
Log-normal ~ unity offset	4177.15	0.2814	0.0076	*δ*	*1*	*fixed*	
Log-normal ~ fitted offset	4165.92	0.370	0.043	δ	11.577	5.428	0.96
Top log-normal	4172.62	0.352	0.035	β	1.891	0.383	0.94
Box-Cox *as 'Power'*	4167.15	0.616	0.171	α	0.521	0.055	-0.99
Box-Cox	4167.15	0.321	0.056	α	0.521	0.055	-0.98
Box-Tukey	4165.94	1.444	0.208	α	0.194	0.048	-0.89
				δ	15.707	4.004	0.31

The log-normal with fitted offset gives a very similar fit to the Box-Tukey (with one less degree of freedom, but no mathematical equivalence is apparent.

The ratios of fitted coefficients to their standard errors, shown in table 4.18, are all substantially significant, apart from the fitted offset δ for the log-normal, which is only just significant. These 't' ratios are much larger for single-parameter models, because two-parameter models have close correlations between their parameters.

These correlations are shown in the last column. They could pose problems in calibration, but none has been apparent. They make single adjustments to one coefficient alone unwise.

4.7.2 Fitted models

Figure 4.22 plots the fitted functions on the linear scale, so the Exponential function appears as a straight line. Because of the balancing factors in trip distribution, the origin of the vertical scale is arbitrary, so individual plots can be moved up or down for comparison with others. The horizontal range corresponds roughly with observed travel costs.

Over most of this range, the non-linear functions are all very similar to the Tanner, which tends to lie between its components, the Exponential and Power. The functions only diverge markedly for short trips, shown in figure 4.43.

Figure 4.42 Fitted non-linear functions

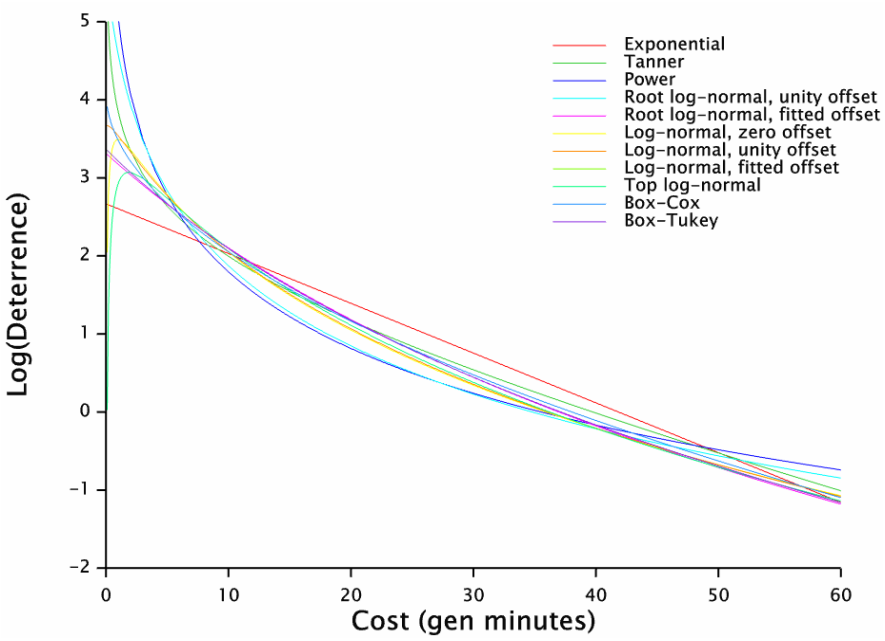


Figure 4.43 Fitted non-linear functions - detail

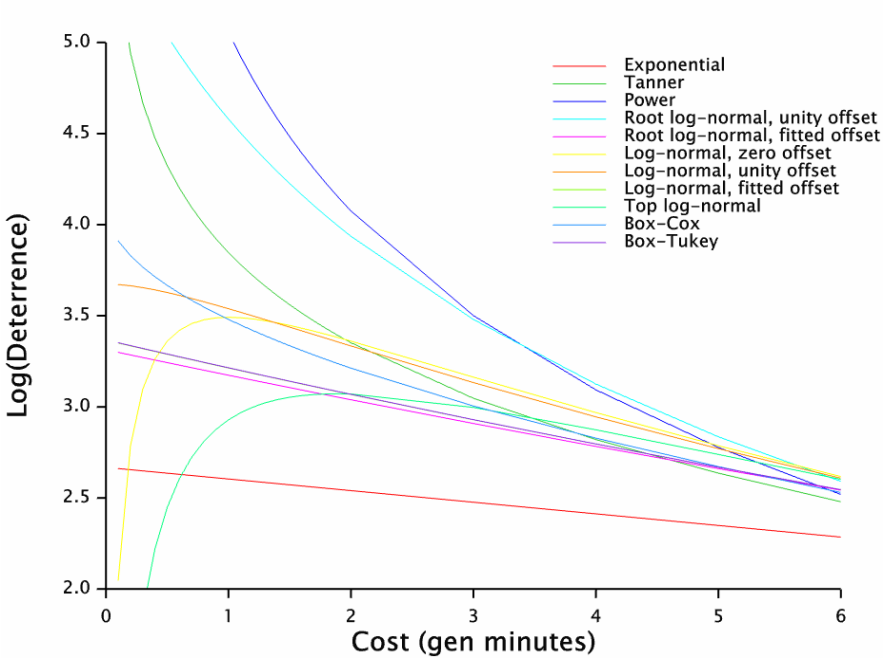
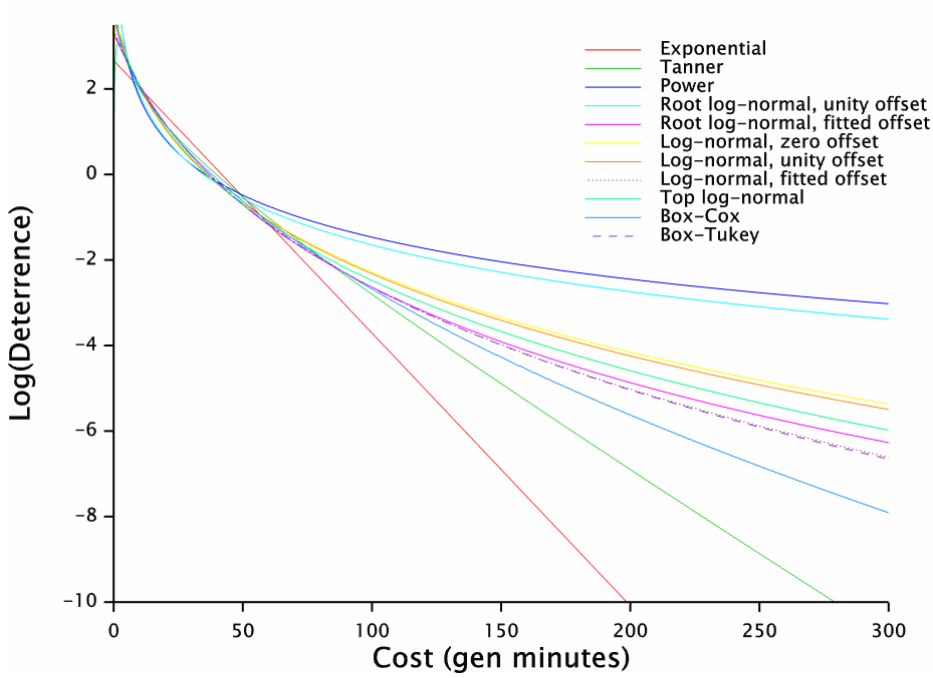


Figure 4.44 Fitted non-linear functions – full range



The log-normal with fitted offset and the Box-Tukey are not distinguishable from each other in these plots; they are just distinguishable from the root log-normal with fitted offset in the detail and full range plots, figures 4.43 and 4.44.

The shortest movements as modelled tend to be in the CBD where zones are smallest, but there are few homes to generate commuter trips in the CBD and generalised costs tend to be increased by parking charges. Hence there may be relatively few observed or modelled trips where functions diverge markedly close to the vertical axis, and many of these will be intrazonal.

Consideration of various intrazonal cost formulations in section 4.11 shows the Power function is very sensitive to them. The Power function increases most markedly towards the vertical axis, but the log-normal and Box-Cox do not do this as much as the Tanner, which appears robust to different formulations of intrazonal costs. In none of these did the Tanner 'turn over' and start to descend near the axis, as the top log-normal and log-normal with zero offset do. Although the 'turn over' may be so close to the axis as to have no practical effect, it is intuitively implausible and an undesirable feature of a function in the absence of demonstrable need.

Non-linear functions generally show a greater probability of very long trips than the Tanner.

4.7.3 Alternative formulations

The non-linear coefficient of the Tanner function, δ in table 4.18 is 18.191. This corresponds with the ratio of the coefficients from the linear model 0.6620/0.0364 =18.203, which is used to calculate Tannerised costs in section 4.2.4. The estimated standard error for the ratio is 3.61, slightly less than the 4.16 shown for the non-linear coefficient.

The top log-normal function $-\lambda(\log(\text{cost}/\beta))^2$

$= -\lambda(\log(\text{cost})-\log(\beta))^2$

$$= -\lambda(\log(cost))^2 + 2\lambda\log(cost)\log(\beta) - \lambda(\log(\beta))^2$$

$$= a(\log(cost))^2 + b\log(cost) + c$$

This a quadratic, or polynomial of order 2, in log(cost), which can be fitted as a linear model with terms in log(cost) and $(\log(cost))^2$.

Table 4.19 Coefficients of top log-normal function fitted by alternative formulations

Non-linear			Quadratic		
Coefficient	**Fitted value**	**se**	**Coefficient**	**Fitted value**	**se**
$-\lambda$	-0.352	0.035	a	-0.353	0.030
$2\lambda\log(\beta)$	0.449	~	b	0.449	0.164
β	1.891	0.383	exp(-b/2a)	1.892	0.341

Similarly, the Box-Cox function

$$\lambda(cost^{\alpha}-1)/\alpha$$

can be re-written as

$$(\lambda/\alpha)cost^{\alpha} - \lambda/\alpha$$

which is equivalent to the Power transformation

$$\lambda' cost^{\alpha}$$

that appears as item G in Daly's list of cost damping mechanisms, plus a constant absorbed into the balancing factors. Again the equivalence can be checked in the fitted coefficients in table 4.18.

The residual deviances for alternative formulations agree to better than three decimal places. Together with the close correspondence in fitted coefficients, this gives confidence in the algorithms.

4.7.4 Statistics for model comparison

With the fitting of variable offsets in the log-normal and Box-Cox (giving the Box-Tukey) and the introduction of the root log-normal, many of the models are closely related. A fully fitted model can be reduced to another model form by fixing one of the non-linear coefficients, usually to zero or unity. These relationships are shown by lines joining the models in figures 4.40 and 4.41 and listed in table 4.20. In simple regression, setting a coefficient to zero is equivalent to dropping the corresponding term from the full model, giving a nested model.

The significance of this reduction can be tested either by comparing the change from the fitted coefficient with its standard error, or by comparing the deviances of the two models. In linear models with normal errors, the change in deviance matches the square of the t statistic (estimate/standard error) for the fitted coefficient. This has been found to be a close approximation for many nested GLMs (section 4.3.4), though not for a simple, non-nested comparison of Power and Exponential functions (table 4.2). However, for comparison between the non-linear GLMs listed in table 4.20, the alternative statistics shown in the final two columns are markedly different. Neither statistic is consistently the larger and the difference can be more than an order of magnitude.

Table 4.20 Statistics comparing non-linear models

Fully fitted model	Fixed coefficient	Reduced model	Change in deviance	(t statistic)^2
Box-Tukey	α=0	Root log-normal, δ fitted	0.10	16.34
Box-Tukey	α=1	Exponential	83.76	281.96
Box-Tukey	δ=0	Box-Cox	1.21	15.39
Box-Cox	α=0	Power	187.35	89.73
Box-Cox	α=1	Exponential	82.55	75.85
Log-normal	δ=0	~	14.71	4.55
Log-normal	δ=1	~	11.23	3.80
Root log-normal	δ=0	~ ; or Power	188.46	13.52
Root log-normal	δ=1	~	128.32	12.54
Top log-normal	β=1	Log-normal	8.01	5.41
Tanner (non-linear)	δ=0	Exponential	175.31	19.13
Tanner (non-linear)	$\delta\rightarrow\infty$	Power	70.47	$\rightarrow\infty$
Linear formulations	Omitted term			
Tanner	log(cost)	Exponential	175.31	81.09
Tanner	cost	Power	70.47	109.06
Top log-normal	log(cost)	Log-normal, δ = 0; or log^2(cost)	8.01	7.50
Top log-normal	log^2(cost)	Power	181.88	138.45

~ Same terminology as fully fitted model, but with offset coefficient δ fixed; this offset is fitted in the full model.

As an extreme case, shown as the last non-linear formulation in table 4.20, the Tanner function can be seen as tending to the Power function as the non-linear coefficient δ becomes large, weighting the log(cost) component relative to the cost component. As δ tends to infinity, so must its t statistic.

The standard error reflects the sensitivity of the deviance to the estimate of the fitted coefficient. The calculation of the standard error may involve some approximation in a non-linear GLM and applies at the best-fit estimate of the coefficient. Using the standard error to extrapolate the change in deviance to a value of zero or unity by the using the t statistic is clearly unsafe in the cases considered here. Choice between model forms should be made on differences in deviances, which can always be calculated from fitted and observed data.

4.7.5 Model fitting in Genstat

The non-linear function *f*() was introduced as the `CALCULATION` option of the `FIT` directive. The function F had been declared previously as an expression, eg

```
EXPRESSION BoxTukey; !e(F=((Cost+D)**A-1)/A)
```

for the Box-Tukey. The non-linear coefficients were given plausible starting values in the `INITIAL` parameter of the `RCYCLE` directives, from values found in the literature review or by consideration of similar models. For example, the coefficient δ in the non-linear form of the Tanner function can be calculated from the ratio of the coefficients from the linear form as 18.2; the initial value was set to 20. The functions were plotted with these initial values to check they gave a reasonable form.

A starting set of fitted values was introduced in the `FITTEDVALUES` option of `RCYCLE`. This was taken from the fitted values of a (linear) Tanner model for all forms of non-linear model. `RCYCLE` was also used

to increase the number of iterations to MAXCYCLE=30, as this sometimes gave convergence which was not achieved by the lower defaults. Other settings for the algorithms, such as METHOD, STEPLENGTH and TOLERANCE, were left at their default values. The RCYCLE directive had to follow the MODEL directive.

The EVA and double forms of model failed to converge. These all have two non-linear coefficients to be fitted. Only the Box-Tukey converges with two non-linear coefficients. From the last column of table 4.18 its coefficient δ has an unusually low correlation of 0.31 with the linear coefficient λ, and the correlation of α with λ is lower than in models with single non-linear coefficients, so it may be unusually well parameterised for fitting. Given the good fit of some models with single non-linear coefficients, or even their reduced linear forms, more complex models may offer little improvement in fit.

Other approaches considered were:

- fitting one non-linear coefficient while fixing the others and iterating between them
- finding good initial values by fitting to the fitted values of other well-fitting models
- finding good initial values by fitting to a trip length distribution aggregated by cost, or the residuals from a trial distribution, as in the KALIBRI algorithm (section 2.6.1.4)
- checking the algorithms can recover non-linear coefficients from a distribution synthesised from them.

The double Exponential and Power (TModel) functions and the Schiller reduced form of the EVA2 function all require the linear coefficient λ to be fixed at unity, to avoid fitting a power other than unity over the whole function. An attempt was made to formulate this by introducing the non-linear function $f()$ as the OFFSET option in the MODEL directive, but no convergence was achieved, even for simple, established models such as the Exponential when formulated in this way.

Genstat offers standard non-linear curves for regression (FITCURVE), but none of these appear to coincide with deterrence functions found in the literature review. Genstat can undertake general model optimisation (eg FITNONLINEAR). However, both of these would probably lose the advantage of the GLM algorithms for fitting the large number of trip end balancing factors as log-linear components.

Fitting and interpreting non-linear models is something of an art and no great effort was made to refine it.

4.7.6 Preferred forms

Many of the non-linear models and even their reduced linear forms with fixed offsets fit well.

The root log-normal with fitted offset and the Box-Cox are the models carried forward for further consideration. Both are reduced forms of the Box-Tukey, the only model that could be fitted with two non-linear coefficients. However, the Box-Tukey does not improve significantly on either of them. Neither can be reformulated as linear models.

The root log-normal is a simplification of the log-normal, and is an intermediate, single step from the Power function. The root log-normal is monotonic, whereas the log-normal can have a turning point. The offset fitted to the root log-normal is larger than that for the log-normal. The root log-normal fits slightly less well than the log-normal with fitted offsets and much worse with offsets fixed to unity or zero.

The Box-Cox is a combination of the principal models, the Exponential and Power, and fits even better than their linear combination in the Tanner function.

4.8 Statistical measures of fit

4.8.1 Residual deviances

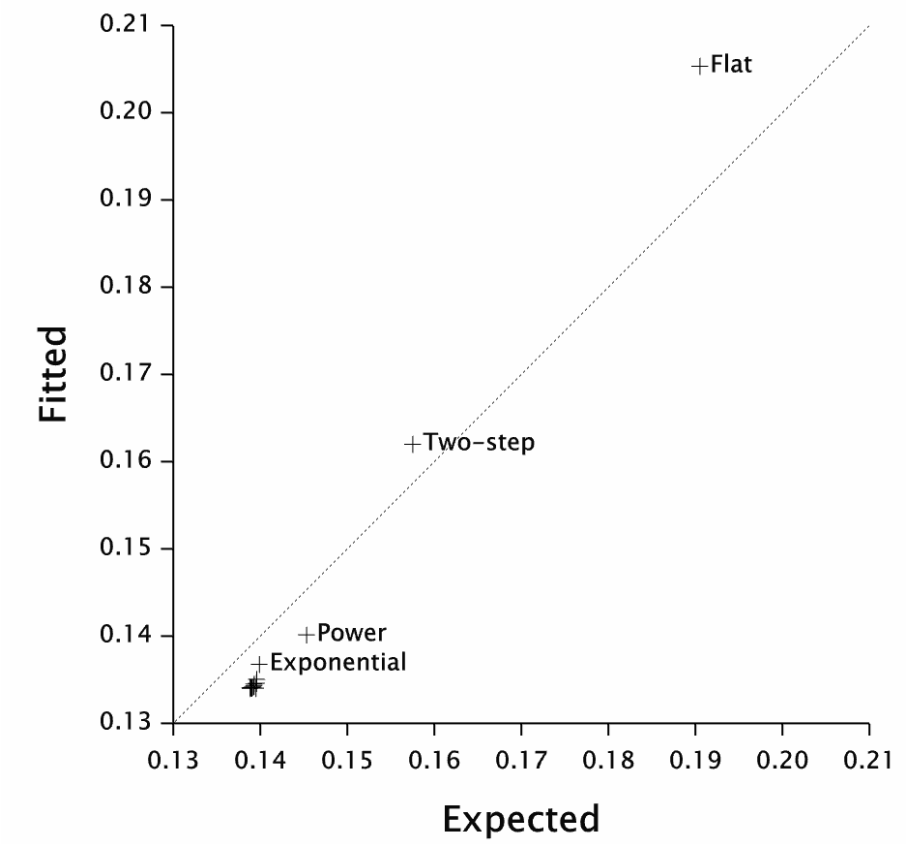
The goodness of fit of a model can be checked against the residual deviances. In a well-fitting Poisson model with expectations larger than unity, the residual deviance approximates to a chi-squared distribution. Section 3.4 shows this is not the case for sparse Poisson models, with expectations less than unity. However, an expected residual deviance and its variance can be calculated from the fitted values of a model. This has been done for a selection of the deterrence functions considered in the preceding sections. The expected and fitted residual deviances are listed in table 4.2.1 and plotted in figure 4.45.

Table 4.21 Mean residual deviances

Deterrence function		Expected deviance		Fitted deviance	
Model	**df**		**SE**		**t statistic**
Flat	0	0.1905	0.0043	0.2053	3.48
Exponential	1	0.1399	0.0035	0.1368	-0.89
Power	1	0.1453	0.0038	0.1401	-1.37
Tanner	2	0.1393	0.0036	0.1345	-1.34
Empirical two step	1	0.1575	0.0039	0.1620	1.14
Empirical two slope	2	0.1396	0.0036	0.1351	-1.27
Empirical five slope	5	0.1388	0.0036	0.1340	-1.35
Geographic segmentation	10	0.1396	0.0036	0.1346	-1.39
Tanner spline O(2)	3	0.1392	0.0036	0.1342	-1.40
Tanner spline O(3)	4	0.1390	0.0036	0.1340	-1.39
Spline log O(4)	4	0.1395	0.0036	0.1340	-1.51
Tanner polynomial O(2)	3	0.1388	0.0036	0.1341	-1.33
Polynomial log O(2)	2	0.1393	0.0036	0.1343	-1.38
Polynomial log O(3)	3	0.1389	0.0036	0.1341	-1.34
Root log-normal, fitted offset	2	0.1389	0.0036	0.1341	-1.33
Box-Cox	2	0.1390	0.0036	0.1341	-1.36
Box-Tukey	3	0.1389	0.0036	0.1341	-1.34

The expected deviances depend upon the expected values in each cell and thus vary between models. Table 4.21 and figure 4.45 show a close relationship between the fitted and expected residual deviances. Only for the flat model, with no cost deterrence at all, is the fitted residual deviance significantly higher than expected. Even the crude empirical two-step model, simply treating trips as either long or short, shows no significant lack of fit and for all other deterrence functions the mean fitted deviance is less than expected.

Figure 4.45 Mean residual deviances



Comparing fitted with expected residual deviances for a single model is ineffective as a measure of model fit. Although the fitted residual may be similar to the expectation for that model, there may be other models with lower expectations that show the original fitted residuals in a much worse light.

There is no obvious reason why models which fit better due to the addition of terms should have lower expected residual deviances, but this appears to be the general case. In one case not shown here there is a slight increase in the mean expected deviance with a higher order spline, but degrees of freedom are approximate in splines. The mean fitted residual deviances are calculated from the residual degrees of freedom, so the mean fitted deviances can increase with the addition of an insignificant term.

4.8.1.1 Influence of weighting

Although changes in a single weight applied to all the data do not affect the systematic model, they do affect the expected residual deviances in a complex manner. They factor the expectations on the true Poisson scale of events, giving the distributions of effective counts in figure 4.7; a change in weighting produces a shift along its horizontal axis. The expected total residual deviance is found by summing the product of deviance from figure 3.6 and the number of cells with that effective count from figure 4.7. Weighting affects the relative horizontal position of these curves.

Despite this complex relationship between weight and residual deviance, the ratio of fitted to expected deviances does reduce with the weights considered in table 4.22.

Table 4.22 Residual deviances from Tanner deterrence function

Weight	Source	Expected	Fitted	Ratio
1/40	WTSM; all purposes and modes, including roadside and extra rail interviews	0.287	0.531	1.85
1/60.2	HBW car trips; simple expansion	0.236	0.353	1.49
1/74.7	HBW car trips; with allowance for unequal sampling	0.212	0.284	1.34
1/105.3	Workplaces; simple expansion	0.176	0.202	1.15
1/157.9	Workplaces; with allowance for unequal sampling	0.139	0.135	0.96
1/191.6	Workplaces; erroneous bias by trip frequency	0.124	0.111	0.89

The same trend in ratio appears for other deterrence functions in table 4.23. Residual deviances appear to reflect the error model (weighting) as much as the systematic model (deterrence function).

Table 4.23 Ratios of fitted to expected deviance

Weight	Deterrence function				
	Flat	Two-step	Exponential	Power	Tanner
1/40	1.95	1.91	1.89	1.80	1.85
1/60.2	1.59	1.56	1.52	1.46	1.49
1/74.7	1.45	1.41	1.37	1.32	1.34
1/105.3	1.25	1.21	1.16	1.14	1.15
1/157.9	1.08	1.03	0.98	0.96	0.97
1/191.6	1.01	0.96	0.91	0.90	0.89

The low ratios for the Power function are probably due to a higher expected deviance arising from fewer very sparse cells. In other measures the Power is inferior to the Exponential and Tanner.

4.8.2 Change of deviance with overfitting

Although sparsity changes the properties of residual deviances, section 3.6 demonstrates that changes in deviance are more robust. They change simply in proportion to an overall weight and approximate to the χ^2 distribution when a random (uncorrelated) term is added to a model.

Most deterrence functions show no statistically significant improvement after the first two or three terms, but several models have been fitted with many more terms, in part due to their apparent significance under higher weights in early runs and to explore the limits of GLMs. The overfitted terms may be expected to act as random terms and reduce the deviance by about 1 per degree of freedom. Table 4.24 shows the change of deviance with overfitting of several deterrence functions.

Table 4.24 Change in deviance with overfitting

Deterrence function	Initial model		Added terms		
	deterrence fn	residual	change in		Deviance
	df	deviance	df	deviance	df
Empirical					
Slope	5	4155.6	5	3.69	0.74
Slope and step	9	4150.9	10	4.69	0.47
Splines					
Spline	5	4152.2	5	7.34	1.47
Tanner spline	6	4155.5	5	3.08	0.62
Tanner spline(log)	6	4151.6	5	5.44	1.09
Spline(log)	5	4152.4	5	4.83	0.97
Tannerised cost	5	4155.3	5	3.36	0.67
Polynomials					
Polynomial	5	4153.5	5	6.39	1.28
Tanner polynomial	6	4156.9	5	2.89	0.58
Tanner polynomial(log)	6	4148.9	5	7.24	1.45
Polynomial(log)	5	4151.0	5	6.15	1.23
Tannerised cost	5	4157.1	5	3.27	0.65
Overall			**65**	**58.35**	**0.90**

The overall change in deviance per degree of freedom is 0.90, quite close to unity. The range of values for individual deterrence functions is quite wide. Taking a conservative estimate that the range arises from a common randomness in the data, a χ^2 distribution with five degrees of freedom gives a 95% confidence interval from 0.167 to 2.57. Assuming that the different deterrence functions give independent results, the range with 65df is 0.69 to 1.37. The result of 0.90 is well within these ranges, but they do not provide a particularly tight check on the weighting scale.

Despite this, residual deviances do seem to reflect as much on the adequacy of the error model represented by the weighting scale as they do on the fit of the systematic model.

4.8.3 Systematic models

Table 4.25 summarises the extent to which different deterrence functions explain the systematic component of deviance. Deterrence functions are set out down the table in increasing complexity and fit to show the reduction in deviance from one form to the next, in single steps between nested models wherever possible. Although a single common weighting factor can affect the scale and significance of the differences between models, their order and relative differences are unaffected.

Table 4.25 Fit of deterrence functions

Model	Residual		Change			Sample size for detection	
	df	Deviance	df	Deviance	%	Households	Trips
Flat (upper bound)	**31,073**	**6380.4**		**0**	**0%**		
2 step empirical	31,072	5033.7	1	1346.8	60.8%	30	36
=> empirical 2-step, common slope	*31,071*	*4229.8*	*1*	*803.9*	*36.3%*	*51*	*61*
Power	31,072	4354.5	1	2025.9	91.4%	20	24
=> Tanner	*31,071*	*4179.2*	*1*	*175.3*	*7.9%*	*232*	*278*
=> polynomial log O(2)	*31,071*	*4172.6*	*1*	*181.9*	*8.2%*	*223*	*268*
=> polynomial log O(2->3)	31,070	4165.4	*1*	*7.2*	*0.3%*	*5611*	*6732*
=> root log-normal, fitted offset	31,071	4166.0	*1*	*188.4*	*8.5%*	*188*	*215*
=> Box-Cox	*31,071*	*4167.2*	*1*	*187.3*	*8.5%*	*187*	*217*
Exponential	31,072	4249.7	1	2130.8	96.2%	19	22
=> Tanner	*31,071*	*4179.2*	*1*	*70.5*	*3.2%*	*576*	*691*
=> empirical 2-step, common slope	*31,071*	*4229.8*	*1*	*19.9*	*0.9%*	*2042*	*2450*
=> empirical joined slopes (2)	*31,071*	*4196.5*	*1*	*53.1*	*2.4%*	*764*	*917*
=> empirical joined slopes (2->5)	*31,068*	*4163.3*	*3*	*33.3*	*1.5%*	*3663**	*4395**
=> geographic segmentation, K&L	*31,064*	*4180.5*	*8*	*69.1*	*3.1%*	*4698**	*5637**
=> Box-Cox	*31,071*	*4167.2*	*1*	*82.5*	*3.7%*	*83*	*492*
Tanner	31,071	4179.2	2	2201.2	99.4%	37*	44*
=> spline O(2)	*31,070*	*4168.6*	*1*	*10.6*	*0.5%*	*3830*	*4595*
=> spline O(2->3)	*31,069*	*4164.1*	*1*	*4.5*	*0.2%*	*8993*	*10,790*
=> polynomial O(2)	*31,070*	*4165.4*	*1*	*13.8*	*0.6%*	*2945*	*3534*
=> the third factor	*31,070*	*4165*	*1*	*14.2*	*0.6%*	*2862*	*3434*
Nominal lower bound	**31,070**	**4165**	3	2215.4	**100%**	55*	66*

Regular type: prime model, showing changes from the initial flat model, introducing a cost effect

=> Italics, incremental change from model above with one less indent, refining the cost effect

Sample sizes are scaled from WTSM samples of 2538 households and 3045 trips to give a change in deviance of 16/df.

* overestimated sample sizes for multiple degrees of freedom

The scope for fitting an effect of cost is taken between the flat model at the top of the table and a nominal lower bound at a residual deviance of 4165 and 31,070 degrees of freedom, by inspection of the overfitted splines and polynomials.

A hierarchy of models is given in the first column. Prime models are shown in regular type on the left of the column. Their deviance changes are from the flat model.

Other models are shown in italic type, indented to show their level in the hierarchy. Their deviance changes are calculated from the prime models or intermediate models above them in the list and hierarchy, breaking their development into steps of one degree of freedom where possible. Because the Tanner and Box-Cox can be developed from both the Power and Exponential models, they appear as increments to both as well as the Tanner appearing as a prime model in its own right. The two-step empirical model with common slope is similarly developed from both the two-step and the Exponential model. A nominal 'third factor' is postulated to fill the gap between the Tanner and the lower bound.

The residual deviance column confirms that 4165 is a reasonable limit for the lower bound, if the two slightly lower deviances are viewed as containing an element of overfitting or random effects.

The percentage column compares changes in deviance with the range between the flat model and the lower bound. Even the coarse two-step empirical model accounts for more than half the range, while the simple Exponential explains over 96% of this systematic component. This leaves relatively little room for improvement. The Tanner can account for much of this, leaving only 0.6% from the lower bound, though this figure will be sensitive to the actual value of the limit that cost effects can explain.

All these changes in deviance are significant against a 95th percentile of 3.84 for χ^2_1, and the introduction of the prime effects of cost is hugely significant.

4.8.4 Sample sizes

The high significances are in part due to the large WTSM survey. Sample sizes that are just sufficient to reliably find significant models or increments are shown in the last two columns of table 4.25. They are calculated simply by factoring down WTSM sample sizes of 2538 households and 3045 HBW trips by car to give a change in deviance of 16 per df. This target, rather than 3.84, allows not only for a 5% significance test, but also for a 95+% chance of detection (power of test) at this significance level. The target is taken from 2σ for significance plus 2σ for power, all squared, giving $16\sigma^2$ where σ is the standard error. This may be conservative, exaggerating sample sizes. It is based on a single degree of freedom and so is conservative over multiple degrees of freedom. The calculation does not allow for any effects of increasing sparsity on the sampling error or other departures from large-number theory.

Commuting by car is likely to have the largest sample of trips per household of any purpose and mode. Sample sizes are therefore expressed in trips as well as households. Distributions of other purposes and modes may have less well-defined cost deterrence effects, which would also require an increase in sample size to detect and calibrate them.

At first sight, the sample sizes seem small with as few as 20 for prime models. On consideration, cost deterrence is simply people's reluctance to travel further than they need; it is a clear, credible effect and it should not be difficult to detect. The complexities of trip distribution need not obscure it; indeed, efficient calibration should not do so.

The weighting applied here allows for a lack of independence in trips to and from a person's workplace. Samples of independent trips from a roadside, station or on-vehicle surveys could be smaller.

A consequence of a powerful sample is that cost deterrence can appear significant in small subsets of the data, such as Peugeot drivers, shop assistants, or the over 60s. There is a risk of over-refining a model into such subsets, which produce individually significant cost coefficients, although the distinctions between them are much less significant. This is a known problem in formulating complex models and is addressed by testing the differences.

Empirical models comprising subsets defined by cost ranges provide a case in point. Table 4.12 shows the five-band model in the middle columns. The absolute values of slopes at the bottom of the table are effectively cost coefficients for the individual subsets. All are highly significant, taking a critical value of 2 for the t statistics. However, only half of the differences between them (shown in the middle rows of the table) are significant and by a much smaller margin.

The significance of cost coefficients for individual cost bands is relatively low compared with other subsets of similar sample size, because their range of costs is limited by definition. Geographic segments as used in the WTSM are restricted to a lesser extent. Figure 4.21 shows the spread of costs for each segment.

Subsets based on socio-economic group, origin or destination, but encompassing a full range of costs, could show similar significance from smaller samples.

The size of the WTSM study area and hence the wide range of costs observed could make the internal household interview dataset more powerful than that for smaller study areas. This would be affected by the treatment of external trips.

Incremental improvements in the cost function generally require at least an order of magnitude more data to achieve significance, reflecting the large proportion of the cost effects captured by the prime models. Thousands of households or trips are needed to improve on the Tanner model. The notable exception is adding a common slope to the poor two-step empirical model.

The sample sizes in table 4.25 are those needed to reliably detect cost deterrent effects or terms in them as just significant. A target change in deviance of 16 is expected to fit coefficients with a standard error of a quarter of their mean (t ratio = 4, =√16). For better accuracies, the target change in deviance – and hence sample size – should increase in proportion to the square of the target t ratio (mean/standard error).

Thus a 5% relative standard error (t=20) might be expected for the cost coefficient in an Exponential model calibrated on a sample of 476 households or 572 trips. This is slightly less than is needed to reliably detect the Tanner's improvement over the Exponential. However, this simple relationship between change in deviance and the relative standard error is not exact in GLMs; in particular, fitted coefficients of cost tend to have larger standard errors than expected from this relationship when a common slope is first introduced into a model (table 4.2).

There is no clear criterion for an adequate calibration, but these calculations suggest that samples in the hundreds may be sufficient rather than in the thousands.

An alternative approach is to consider the sampling error in total travel, which will be replicated in a fitted Exponential distribution model, or one with an equivalent component of cost. The unexpanded sample of car commuting trips in the WTSM has a mean generalised cost of 26 generalised minutes, with a standard deviation of 24. This will require a sample of almost 100 trips to give a 10% relative standard error, or 10,000 for 1%. These samples might be increased for uneven sampling, or applied over all purposes for total travel on the network. This consideration of sampling error applies to observed matrices as much as to synthesised ones and thus omits modelling error.

4.9 Practical measures of fit

4.9.1 Introduction

The development and testing of deterrence functions so far has been based on deviance, a statistical measure of their fit to the household survey data from which they are calibrated. This section considers two more practical measures of the modelled distributions: their fit to observed counts of traffic crossing screenlines, which is an independent measure; and predictions of the use and benefits of schemes, which is the ultimate use of transportation models.

The various deterrence functions have been calibrated on the matrix observed from the HIS for commuting trips internal to the study area only. Deterrence functions have been chosen as effective examples from each form, except for the empirical two-step which is included for its coarseness. The empirical five-slope and geographic segmentation take more degrees of freedom, which may help their relative performance.

4.9.2 Screenlines

See appendix C for further details of screenlines.

The central screenline forms a cordon around the central area, so there are many through movements crossing it twice, unlike the other screenlines. Directions 'in' and 'out' are toward and away from the centre of Wellington.

Table 4.26 shows estimated crossings of four screenlines, in each direction for three periods of the day. The first line for each screenline gives the crossing counts; since trip purpose cannot be distinguished from vehicle counts, the original counts have been factored to HBW using the WTSM base model. The external trips observed at roadside surveys on the study area boundary have been subtracted from the traffic counts. All flows are shown as hourly rates.

The second line for each screenline shows the crossings assigned from the matrix observed in the HIS. This 24-hour production–attraction matrix is factored by WTSM direction and period models before assignment. Models calibrated from the observed matrix are factored in the same way and the same assignment is applied to find their crossings of the screenlines. The adjustment of counts and assignment of matrices is described in section 8.6.3.

Table 4.26 Screenline crossings

Screenline	Model	AM in	AM out	IP in	IP out	PM in	PM out
Central	**Screenline count**	**9730**	**2500**	**1076**	**827**	**1910**	**6069**
	HIS observed matrix	8792	2143	934	687	1655	5506
	Flat	14124	7311	1655	1354	5208	9257
	Two step	9951	3233	1092	820	2424	6371
	Power	9272	2557	1004	741	1962	5887
	Exponential	9063	2380	995	749	1850	5718
	Tanner	8990	2300	977	726	1791	5676
	Five slope	8835	2153	957	709	1693	5567
	Geographic segmentation	8916	2229	970	722	1746	5623
	Tanner spline O(2)	8927	2242	969	719	1753	5633
	Polynomial log O(3)	8870	2185	961	712	1715	5595
	Root log-normal with offset	8826	2142	955	706	1686	5564
	Box-Cox	8894	2207	964	714	1730	5611
Radial	**Screenline count**	**3891**	**2035**	**480**	**350**	**1474**	**2551**
	HIS observed matrix	4148	1963	464	352	1370	2709
	Flat	9003	6813	1162	1047	4715	6062
	Two step	5358	3166	649	533	2223	3574
	Power	4810	2621	557	443	1831	3179
	Exponential	4386	2201	489	377	1533	2873
	Tanner	4414	2227	496	383	1554	2898
	Five slope	4239	2052	470	358	1434	2777
	Geographic segmentation	4268	2082	479	366	1457	2799
	Tanner spline O(2)	4347	2161	486	373	1509	2851

Screenline	Model	AM in	AM out	IP in	IP out	PM in	PM out
	Polynomial log O(3)	4294	2107	479	366	1473	2816
	Root log normal with offset	4248	2061	473	360	1441	2784
	Box-Cox	4323	2136	483	370	1492	2836
Regional	**Screenline count**	**697**	**119**	**73**	**61**	**100**	**519**
	HIS observed matrix	632	156	69	57	119	403
	Flat	2156	1673	328	309	1113	1407
	Two step	1507	1019	221	199	682	981
	Power	938	453	126	105	310	606
	Exponential	547	71	58	46	61	345
	Tanner	609	129	69	54	99	387
	Five slope	647	165	75	58	122	414
	Geographic segmentation	612	133	70	55	102	389
	Tanner spline O(2)	617	136	70	54	103	393
	Polynomial log O(3)	620	140	71	55	106	396
	Root log normal with offset	627	145	72	55	109	400
	Box-Cox	621	140	71	55	106	396
Rimutaka	**Screenline count**	**401**	**68**	**37**	**37**	**72**	**255**
	HIS observed matrix	525	105	59	58	82	325
	Flat	2918	2473	457	435	1639	1913
	Two step	1923	1479	297	275	981	1254
	Power	886	452	123	110	308	568
	Exponential	532	111	62	61	85	329
	Tanner	541	117	63	59	89	337
	Five slope	575	149	69	64	110	360
	Geographic segmentation	553	130	66	62	97	345
	Tanner spline O(2)	556	132	66	62	99	347
	Polynomial log O(3)	566	142	68	63	105	354
	Root log normal with offset	583	158	71	65	115	365
	Box-Cox	561	136	67	62	101	350

Table 4.27 summarises differences across all screens, periods and directions, taking either the observed matrix from household travel surveys or the traffic counts as base. Percentages are between totals from table 4.26. GEH statistics, which are less sensitive to smaller values, are the root mean square of GEH values for individual entries in table 4.26.

Table 4.27 Fit at screenlines

Model	Percentage		GEH	
	Observed matrix	Screenline count	Observed matrix	Screenline count
Screenline count	6%	~	4.4	~
HIS observed matrix	~	-6%	~	4.4
Flat	154%	139%	46.8	45.9
Two step	51%	42%	23.7	23.9
Power	20%	13%	10.3	11.2
Exponential	5%	-1%	3.4	4.4
Tanner	5%	-1%	2.4	4.3
Five slope	2%	-4%	1.4	4.8
Geographic segmentation	3%	-3%	1.5	4.3
Tanner spline O(2)	4%	-2%	1.9	4.4
Polynomial log O(3)	3%	-3%	1.6	4.6
Root log normal with offset	2%	-4%	1.6	4.9
Box-Cox	3%	-3%	1.7	4.5

Compared with either screenline counts or the observed matrix, the bad fit of the flat model, which has no cost effect, clearly shows the need for cost deterrence in trip distribution. The poor fit of the coarse two-step model is also apparent and the Power function generally gives a noticeably poorer fit than the rest. These differences tend to be more marked at screenlines further from the centre, probably because the longer trips there are not sufficiently deterred by these functions.

Comparing the better fitting models (Exponential and below in the tables) with the observed matrix from which they are calibrated, the models consistently overestimate crossings of all screenlines except the regional. Here the models underestimate the observed crossings, except for the five-slope model and the inbound interpeak crossings.

These suggest a spatial pattern in the observed matrix that models of cost deterrence cannot fully replicate. Geographic segmentation shows no distinct improvement over its peers. Factoring for direction, period and occupancy are common to both observed and modelled matrices and so are unlikely to be the root cause.

These differences between the observed matrix and the models calibrated from it are generally smaller than those between the observed matrix and the traffic counts. Again, the differences from counts vary between screenlines, with the counts higher at the central screen and lower at the Rimutaka. At the radial screen the counter-peak and IP counts are higher, while at the regional screen the with-peak counts are higher. This suggests some variation in the direction and peaking factors as well as spatially, though it is to be expected that journeys to and from work fit the with-peak counts better.

Overall, the models tend to differ from the observed matrix in the same sense as the counts, reducing the difference between the counts and the models shown by the percentages in table 4.27. The GEH statistics show that differences between counts and the observed matrix continue to reduce the fit of models to the counts. However, both percentages and GEH show the Exponential and Tanner models fitting as well to screen counts as more advanced models; the Power fits best at the central screen.

Values from the observed matrix are themselves subject to sampling error, since only a fraction of the trips recorded in household interviews cross any particular screenline. While the percentages suggest that synthetic matrices may gain in accuracy from other recorded trips, the GEH shows no difference in precision.

There are overall signs that more advanced deterrence functions fit the observed matrix better, but advances beyond the Exponential do not improve the fit to screenline counts.

4.9.3 Schemes

See appendix D for further details of schemes.

Table 4.28 shows the users and benefits of three hypothetical schemes. Users are the number of commuter vehicles on a designated link in the scheme, so they are an index rather than a comprehensive total. Benefits are cost savings in vehicles × generalised minutes. The measures are for the AM period only when commuting is most prominent.

Table 4.28 Scheme effects

Model	Central		Radial		Regional	
	users	benefit	users	benefit	users	benefit
HIS observed matrix	3839	9152	9441	17,361	1106	16,792
Flat	11,358	16,416	22,826	34,922	6543	61,812
Two step	5040	10,311	12,399	22,310	4201	40,865
Power	4394	9456	10,672	19,567	2037	24,118
Exponential	3825	8958	8855	17,059	556	13,841
Tanner	3841	8960	9185	17,548	820	15,325
Five slope	3617	8757	8658	16,897	1005	16,294
Geographic segmentation	3783	8928	8922	17,144	820	15,408
Tanner spline O(2)	3750	8872	8997	17,299	872	15,525
Polynomial log O(3)	3663	8799	8872	17,154	896	15,618
Root log normal with offset	3594	8739	8748	17,001	932	15,786
Box-Cox	3704	8838	8958	17,264	890	15,627

The three schemes are located in the same parts of the network as the first three screenlines of the same name. Unlike screenlines, there is no independent benchmark such as traffic counts, so table 4.29 shows the differences of various models from the observed matrix. The figures for the observed matrix itself are estimated standard errors derived from the number of sampled trips assigned to the scheme. These are substantial compared with differences from the better models.

Table 4.29 Scheme effects – differences from observed matrix

Model	Central		Radial		Regional	
	users	benefit	users	benefit	users	benefit
HIS observed matrix – sampling error	7.5%	5.0%	4.7%	4.9%	12.0%	6.5%
Flat	196%	79%	142%	101%	491%	268%
Two step	31%	13%	31%	29%	280%	143%
Power	14%	3%	13%	13%	84%	44%

Model	Central		Radial		Regional	
	users	benefit	users	benefit	users	benefit
Exponential	-0%	-2%	-6%	-2%	-50%	-18%
Tanner	0%	-2%	-3%	1%	-26%	-9%
Five slope	-6%	-4%	-8%	-3%	-9%	-3%
Geographic segmentation	-1%	-2%	-5%	-1%	-26%	-8%
Tanner spline O(2)	-2%	-3%	-5%	-0%	-21%	-8%
Polynomial log O(3)	-5%	-4%	-6%	-1%	-19%	-7%
Root log normal with offset	-6%	-5%	-7%	-2%	-16%	-6%
Box-Cox	-4%	-3%	-5%	-1%	-20%	-7%

As at screenline crossings, the flat, two-step and Power models overestimate traffic. Unlike screenline crossings, the Exponential and other models mostly underestimate the traffic and travel benefits of the schemes. The patterns between models show a broad similarity with the AM inbound crossings of equivalent screenlines after allowing for the generally lower values.

Users appear more sensitive to model specification than benefits. This may be because differences occur in marginal movements that gain relatively little benefit from using the scheme.

Again there are differences between the schemes, in particular the greater underestimation of the regional scheme. This is most marked for the Exponential model. These tendencies can also be seen in regional screenline crossings.

The central and radial schemes show that choice of deterrence function can affect scheme benefits by at least 2%, or considerably more in the case of the regional scheme or the Power function.

These effects might differ in practice with incremental modelling, or the application of matrix estimation as in the WTSM.

4.9.4 Sampling

As with statistical measures of fit, no obvious criterion for sample size appears from these practical measures. They do show that spatial differences, between different schemes or screenlines, are generally at least as great as those between the better-fitting deterrence functions. While these spatial variations remain unexplained, there may be little practical benefit in fitting cost deterrence functions beyond the Exponential. Considering the sample sizes needed to improve on the Exponential in table 4.25, a sample size in the hundreds may be sufficient. However, capturing the spatial variations to the same accuracy in an observed matrix may demand a far larger sample.

4.10 Measures of separation and generalised cost

There are many different measures of cost or more general separation between zones. In highway modelling, these can include junction delays and congestion; public transport involves fares, walking and interchange. This section examines a few of the measures, principally to explore the sensitivity of calibration to them.

In the foregoing work, costs have been the generalised highway time. The relative weights of its four main components, time and distance in the morning and interpeak, can be examined with GLMs.

As a preliminary, basic Exponential models are calibrated against some much simpler measures of separation.

4.10.1 Crow-fly distances

Direct distances between zones are a very simple measure of separation, which can be abstracted easily from a geographic information system (GIS) or coordinates used for model network plotting. They are crude measures since they do not recognise topographical constraints (considerable around Wellington) or differing speeds by road type, let alone congestion effects. On the other hand, they are stable in that they are not affected by congestion or route changes.

Two measures were calculated between:

- traffic centroids used to plot the network
- centres of gravity of the zones, from GIS polygons.

There is no indication that the traffic centroids were located representatively in the WTSM, and the centres of gravity can be far from the centres of population or employment in zones incorporating large remote areas, so these measures are coarse.

Table 4.30 shows calibrations against these measures of cost. The fitted coefficients λ are about twice those fitted on generalised times, which is consistent with the distance element making up about half the generalised time. Because of such differences in units, the accuracies of coefficients are shown as t ratios. These are highly significant, but once again do not match the square root of the change in deviance ($\sqrt{\Delta}$dev), as they would in simple regression. This statistic is derived from the residual deviance, showing the change from the flat model. This difference in deviances is also expressed as a percentage of the systematic deviance between the flat model and a nominal lower bound deviance of 4165; this is the same percentage as is shown in table 4.25.

These deviance measures show that although crow-fly distances capture most of the separation effects, they are markedly poorer at doing so than the full generalised cost.

The difference between the coefficients for the two variants is small, but their deviances suggest that distances between traffic centroids may fit significantly better.

Table 4.30 Fit of measures of separation

Cost formulation	units	λ	t	√Δdev	Deviance	%
Flat model – no cost effects		0	~	0.0	6380.4	0.0
Crow-fly distances between:						
– traffic centroids	km	0.1457	26.4	44.2	4424.0	88.3
– geographic centers of gravity	km	0.1447	26.6	44.0	4445.6	87.3
Intervening opportunities for:						
– employment	zones	0.0274	32.1	43.9	4454.5	86.9
– residence	zones	0.0273	32.5	43.3	4507.8	84.5
Generalised distribution costs:						
from WTSM base synthesis	gen min	0.06377	28.7	46.2	4249.7	96.2
– less CBD parking charges	gen min	0.06377	28.7	46.2	4249.7	96.2
– CBD intrazonal costs adjusted	gen min	0.06378	28.7	46.2	4249.5	96.2
– all intrazonal costs set to zero	gen min	0.06329	28.9	46.3	4239.8	96.6

Cost formulation	units	λ	t	√Δdev	Deviance	%
Assignment costs:						
generalised as for distribution	gen min	0.0642	28.6	46.2	4246.5	96.3
– AM time	min	0.1171	29.9	46.2	4248.2	96.2
– IP time	min	0.1394	29.6	46.3	4238.4	96.7
– AM distance	km	0.1188	27.5	45.9	4276.7	95.0
– IP distance	km	0.1184	27.5	45.8	4278.9	94.9
best generalisation (3 extra df)				46.3	4236.1	96.8

%: percentage of nominal range of reduction in deviance (6380.4 – 4165) from fitting deterrence functions, table 4.25

4.10.2 Intervening opportunities

This form of model considers the number of opportunities closer at hand than a particular attraction as its deterrence function, without considering the amount of separation between them. The model can be calibrated on the ranking of costs rather than actual costs. The resulting coefficients are deterrence per intervening zone rather than per minute or kilometre. Balancing factors still allow for differences in production and attraction trip ends between zones.

Zones were ranked by the distribution cost as used in the main part of this study. Attraction zones were ranked from each production zone as the intuitive ordering of employment opportunities; production zones were ranked from attraction zones as an alternative form for residential opportunities.

The fitted coefficients in table 4.30 show that opportunities become 2.7% less attractive for every zone that is closer. According to the deviance (but not the t ratios), the fit is slightly worse than crow-fly distances. Again, the measure picks up most of the separation effect, but nowhere near as well as the generalised cost.

Deviances suggest that employment opportunity is significantly the better of the two variants. The t ratios beg to differ (suggesting a better fit than the Exponential cost model!) while there is little difference between the fitted coefficients.

These incongruities might be resolved by fitting both measures in a single model. This is the approach developed below for examining the components of generalised cost, but these simple tests on widely different measures of separation are sufficient to show calibration's sensitivity to them.

4.10.3 Generalised distribution costs

Prior to this section, costs have been the generalised highway time from the distribution synthesis stage of the WTSM base model, described in section 4.1.2 and appendix A. The four components of this, time and distance in the AM and IP, are not saved from the distribution stage, but similar sets of costs are kept from the final assignments where they determine routing. The generalised cost for distribution differs from that in the final assignments in several details:

- parking charges – only applied for distribution
- intrazonal costs – left as zero in assignment skims
- lag – costs are damped by averaging in distribution iterations
- load – matrix adjustment is applied to demand for final assignment
- vehicle operating cost – 15c/km for distribution, 7.5c/km perceived for assignment.

To improve the comparison between the distribution and assignment costs, parking charges were removed from the distribution costs and intrazonal costs were set to zero. The results are shown as three stages in table 4.30.

The first stage subtracts parking charges from all attractions in the CBD to which they had been applied, including intrazonals. Because these are trip end charges, they are absorbed in the balancing factors and the results of calibrating an Exponential model are identical. However, subtracting parking charges from intrazonal costs makes them negative, since they are processed after parking charges are added in the WTSM (appendix A). This means Power or Tanner models, whose calibration might be changed, cannot be fitted.

The second stage re-calculates the intrazonal costs for the CBD zones using the WTSM formulation. This affects the calibration only very slightly, probably because the CBD zones have very few residential productions.

The third stage sets all intrazonals to zero. The change in the Exponential coefficient is small, but there is an improvement in the deviance that could be significant. Because the WTSM formulation for intrazonal costs is non-linear using minimums, it would be difficult to apply while fitting components of generalised cost separately.

The effects of lag and load cannot be determined without further runs of the WTSM base model, saving intermediate workings; their influence in distribution-assignment iterations could be a major study in itself.

4.10.4 Generalised assignment costs

The final assignment saves the four components of generalised cost: time and distance, in the AM and IP periods. The WTSM does not have an evening peak assignment, so the morning represents all travel in the busy periods.

These components can be combined into a generalised cost using the same parameters as used in distribution:

$\alpha = 0.565$	proportion of AM; remainder from IP
$\delta = 0.9268$ minute/km	factor from distance to generalised time
= 15 cents per km vehicle operating cost	
÷ 13.6 cents per minute value of time	
÷ 1.19 vehicle occupancy.	

This is the nearest equivalent to the distribution cost, adjusted for parking and intrazonal costs, that can be derived readily from assignment skims. There is a small change in the fitted coefficient. The increase in deviance could be significant; it is less than the reduction produced by setting intrazonals to zero in the distribution costs.

These are shown in table 4.30 which also shows the fit of components individually and then a model including all four components. Since a separate coefficient is fitted to each component, the ratios between the coefficients can vary. These ratios can re-define generalised cost to give the best fit to the data.

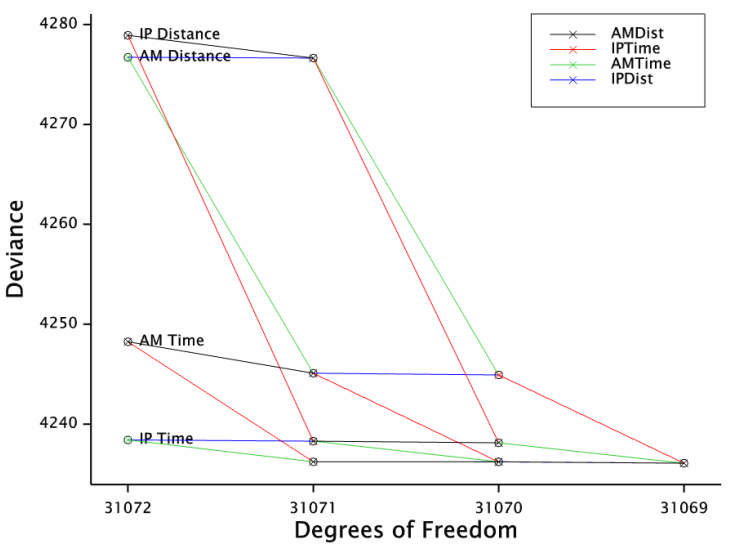
Fitting four coefficients implies an 'interaction term', with separate distance factors for AM and IP, or different mixes of AM and IP for time and distance. Fitting α and δ simultaneously without interaction, so that

$$\text{Cost} = (\alpha\times\text{AMtime} + \alpha\delta\times\text{AMdistance} + (1-\alpha)\times\text{IPtime} + (1-\alpha)\delta\times\text{IPdistance})$$

falls outside the linear form of GLMs. For this reason, there is no single cost coefficient λ to show at the bottom of table 4.30; various combinations and ratios are shown in the following tables.

Figure 4.46 shows the build-up of the full model from its four components.

Figure 4.46 Fit of assignment cost components



Each of the four components can singly capture most distribution effects of the generalised costs.

Times fit better than distances, falling in the same range of deviance as generalised costs. Distances do not fit so well, but far better than crow-fly distances. The coefficients are lower than for crow-fly distances, consistent with longer distances via the network.

More surprisingly, the IP time gives a better fit than the AM one. This could reflect the difficulties of estimating journey times in congested conditions, for both commuters and modellers.

The components are naturally highly correlated, so additional components can give relatively little further improvement in fit. However, some still appear highly significant. Plots are shown only for rational pairings of components, with either period (AM or IP) or dimension (time or distance) in common. These all have 31,071 degrees of freedom. The three-component models (31,070 df) are unbalanced, and are only included for completeness, showing the effects of individual components in adjacent models.

Measures of significance for single components, rational pairings and the full four-component model are shown in table 4.31. Each quartering of the table gives the significance of one component:

- singly (as listed at the bottom of table 4.30), at the outer corner of the table
- paired with another component, on the side adjacent to the other component
- in the full four-component model, in the middle of the table.

Table 4.31 Fit of generalised cost components

Component	Morning AM						Interpeak IP					
	Coef.	t	√Δdev	Coef.	t	√Δdev	Coef.	t	√Δdev	Coef.	t	√Δdev
Time (minutes)	0.117	29.86	46.18	0.035	1.48	1.48	0.098	3.43	3.47	0.139	29.61	46.28
	0.090	5.61	5.62	0.035	1.44	1.43	0.096	2.97	2.97	0.132	6.34	6.37
Distance (km)	0.029	1.77	1.77	0.037	0.38	0.38	-0.035	-0.37	0.37	0.006	0.35	0.35
	0.119	27.47	45.87	0.145	1.51	1.51	-0.026	-0.27	0.27	0.118	27.46	45.84

The coefficients of single component models, which are at the corners of the table, are larger than those for the multiple component models shown in between them. Since the components are broadly similar (for a travel time of one minute per km), the coefficients are divided among them, giving a marked reduction in their individual significances.

In combination with AM distance, IP distance coefficients are perversely negative, but they are not significant. IP distance is only significant as the sole component. Times are generally more significant than distances, with IP the stronger.

The changes in deviance, Δdev, correspond with the slopes in figure 4.46. The $\sqrt{\Delta}$dev and t measures match closely except for the very high significances of single components, where $\sqrt{\Delta}$dev values are higher.

In table 4.32, the coefficients for the logical pairings are converted into factors α and δ for AM and distance components respectively. Standard errors are calculated by Genstat's `RFUNCTION` and used to calculate t statistics for differences from zero, from the WTSM factor, and (for the AM factor α) from unity.

Two intermediate models were also fitted, effectively holding one of the factors α or δ to its value in WTSM distribution while allowing the other to be re-fitted. These are shown at the bottom of each section of table 4.32. Changing the AM-IP mix, α, does not improve on the WTSM factor, but changing proportions of time and distance, δ, does.

Table 4.32 Generalised cost factors

Source	Factor	Standard error	t statistic		
AM proportion	**α**		**≠ 0**	**≠ WTSM**	**≠ 1**
In WTSM distribution:	0.565				
– time	0.267	0.188	1.42	-1.59	-3.90
– distance	1.222	0.813	1.50	0.81	0.27
generalised time, δ = 0.927	0.476	0.387	1.23	-0.23	-1.35
Distance factor	**δ**	**min/km**			
In WTSM distribution:	0.927				
– morning AM	0.321	0.237	1.35	-2.56	
– interpeak IP	0.048	0.143	0.34	-6.15	
averaged, α = 0.565	0.065	0.163	0.40	-5.29	

There is a lack of consistency, or orthogonality. From the t statistics comparing the distance components δ with the WTSM value, there is a good case for a smaller value and (from the comparisons with zero) a reasonable one for none at all. There is no substantial case for changing the proportion of AM, α, from the WTSM or for including either AM or IP distance when the other is present. For a cost based on time alone, there is a significant IP component.

Excluding models of distance alone, differences in generalised cost formulations account for less than 1% of the systematic deviance. They are on the same scale as other 'technical adjustments' to costs, such as setting intrazonals to zero, or the residual differences between distribution and assignment costs, probably due to iteration damping and matrix adjustment. The differences are smaller than many of those between deterrence functions in table 4.25 which failed to show clear improvements in practical terms.

4.10.4.1 Omitting distance

Since vehicle operating costs act through the distance component of generalised cost, omitting the distance component would lose any sensitivity of home-work choice to fuel price.

With deterrence by travel time alone, distribution-assignment iterations may be less stable, since the time component is most directly affected by congestion in assignment. On the other hand, increased weight on the IP with less congestion could improve stability.

4.11 Sensitivity to intrazonal costs

Movements that have their origin and destination within the same zone are known as intrazonal. They do not appear in any part of the model that represents the real network, so assignments are insensitive to their presence. However, they are subtracted from the total generations, so they can affect demand for travel on the model of the real network.

Because intrazonal movements never take a path through the real part of the modelled network, their travel costs are not modelled well, if at all. The costs are small, smaller than interzonal movements and dependent on the granularity in the model imposed by the zoning system.

Although intrazonal trips are poorly modelled, they can make up a substantial proportion of all trips and trip distribution modelling can be quite sensitive to them. To test this sensitivity, a variety of formulations for intrazonal costs have been tried for three simple forms of analytical deterrence function – the Exponential, the Power and the Tanner that combines the first two.

The definition and effects of intrazonal movements depend greatly on the size and shape of zones and their connections into the model network, but these have not been altered from the WTSM model.

4.11.1 Intrazonal cost formulations

WTSM

In the WTSM, intrazonal costs are taken as half of the minimum cost either to or from any other zone, subject to a maximum of five generalised minutes for private vehicles, or 10 minutes for public transport. This ensures that intrazonal movements are always the cheapest to or from a zone, while taking some account of zone size. Zone size varies greatly in the WTSM, from single city blocks in central Wellington to large rural areas of the Wairarapa. WTSM costs are described in section 4.1.2 and appendix A.

Less parking, recalculated

The HBW car model includes parking costs for CBD destinations, which mainly affect the modal split. These are removed to simplify the model, as described in section 4.10.3. Parking costs are first subtracted from all movements to the attraction zones where they were applied, including intrazonals, some of which become negative. Intrazonals are then recalculated to the WTSM formula as described in the preceding paragraph.

The remaining formulations in the following sections all omit parking costs.

Null

A simple approach to intrazonal movements is to exclude them from the distribution model so it is fitted solely to the interzonal movements. This is a reduced dataset compared with all other formulations. In practice, this requires the modelling of interzonal trip generations (productions and attractions) and so just moves the problem to another stage of the modelling.

Zero

Intrazonal costs can be set to zero. This is the result from cost skimming in many software packages. However, Power and Tanner models cannot be fitted because they include the logarithm of cost as a dependent term.

Very small, small, constant, large

This can be overcome by setting all intrazonal costs to a very small value, in this case 0.0001 generalised minutes, which is trivially small compared with any real movement. In the same vein, other values can be applied to all intrazonals:

Very small	0.0001 generalised minutes
Small	0.1 generalised minutes
Constant	1 generalised minute
Large	5 generalised minutes

'Large' corresponds to the cap set in the WTSM formulation and will exceed some interzonal costs, which is implausible.

Zone perimeter

Intrazonal costs may be estimated from the zone's physical size.

The perimeter of the zone is taken from a GIS. On the assumptions that zones are square and intrazonal movements are half the length of one side, the distance is taken as perimeter/8. Perimeters measured from a GIS can be inflated by following boundaries exactly along winding streams or rugged coastlines.

Zone area

Applying the same assumptions as for the perimeter gives a distance of √(area)/2. Where the zoning of a small settled area includes large back-blocks, this will exaggerate movements within the zone.

More sophisticated measures could account for the zones' physical shapes. These have not been tried, mainly for simplicity, but also because traffic generation is often too unevenly distributed to justify any such sophistication, particularly in large rural zones.

Centroid connector lengths – coded and mapped

Centroid connectors are the links in the model between the nominal zone centroids and links representing real roads. They are coded for distances, or their crowfly lengths can be calculated from their coordinates for plotting. The coded distances represent local movement off the 'real' modelled network. No evidence was found that centroid connector mapping was intended to be representative, nor was it obviously unsuitable, eg located in uninhabited parts of zones. Most zones had a single connector, or multiple connectors of the same length – a simple average was taken in other cases.

Centroid connector auto speeds were coded as 40km/h, so distances, including those derived from perimeters and areas, were converted to generalised minutes

$$= \text{minutes} + \text{kilometres} \times 15 \text{ cent/km} / (13.6 \text{ cent/min} \times 1.19 \text{ persons/veh})$$

$$= 1.5 \times \text{kilometres} + 0.927 \times \text{kilometres}$$

$$= 2.427 \times \text{kilometres}$$

These formulations are intended to give a variety of simple measures with some rational basis for testing the sensitivity of trip distribution modelling. They are not intended as candidates for optimal or practical measures.

4.11.1.1 Sample

In the WTSM household survey, 279 intrazonal HBW trips by private vehicle were counted, representing 9% of all HBW private trips, or 1.2% of travel (vehicle-costs) with recalculated WTSM costs for intrazonal movements. Intrazonal movements are a much smaller proportion of travel than of trips because of their short length. Average intrazonal costs for each formulation is given in table 4.33.

Table 4.33 Average intrazonal costs

Intrazonal cost formulation	Average intrazonal cost (generalised minutes)	
	Simple average	Trip-weighted average
WTSM	2.292	3.155
– less parking charges	1.687	3.155
– and recalculated intrazonal costs	2.233	3.155
Null (excludes intrazonals)	~	~
Zero	0	0
Very small	0.0001	0.0001
Small	0.1	0.1
Constant	1	1
Big	5	5
Perimeter	5.422	7.766
Area	3.595	5.168
Coded length of centroid connector	1.575	2.659
Map length of centroid connector	1.448	2.708

Trip weighted costs are not affected in recalculating the WTSM formulation because no intrazonal trips were observed in the CBD zones where parking charges were applied. The change in the simple average is also small after recalculation.

Simple and trip weighted averages are the same for zero, very small, small, constant and big formulations because the same intrazonal cost is applied to every zone.

The high values for perimeter and area formulations could indicate that costs within some zones are higher than those to come or go outside them, but they could just be derived from large zones where high intrazonal costs are appropriate. The differences between perimeter and area demonstrate that zones are not square, as assumed in their formulations. The higher averages under trip weighting show the natural tendency to observe more intrazonal trips within larger zones.

The WTSM formulations fall broadly in the middle of the range of formulations.

4.11.2 Fitted models

Table 4.34 shows the coefficients fitted to the various formulations in three forms of deterrence function. The coefficients of cost and its logarithm are given for the Exponential and Power functions respectively. The Tanner function incorporates both these terms and the ratio between their coefficients is also given.

Table 4.34 Fitted coefficients for alternative intrazonal cost formulations

Intrazonal cost formulation	Exponential		Tanner						Power	
	Cost		Cost		Ratio		Log(cost)		Log(cost)	
	Coef	t	Coef	t	Coef	t	Coef	t	Coef	t
WTSM	0.06377	28.68	0.0364	10.44	18.2	5.04	0.6620	9.01	1.416	43.29
– less parking	0.06377	28.68								
– recalculated	0.06378	28.68	0.0364	10.25	17.7	4.90	0.6456	8.71	1.398	43.08
Null	*0.06045*	*26.53*	*0.0331*	*8.21*	*22.0*	*4.00*	*0.7302*	*7.32*	*1.542*	*36.57*

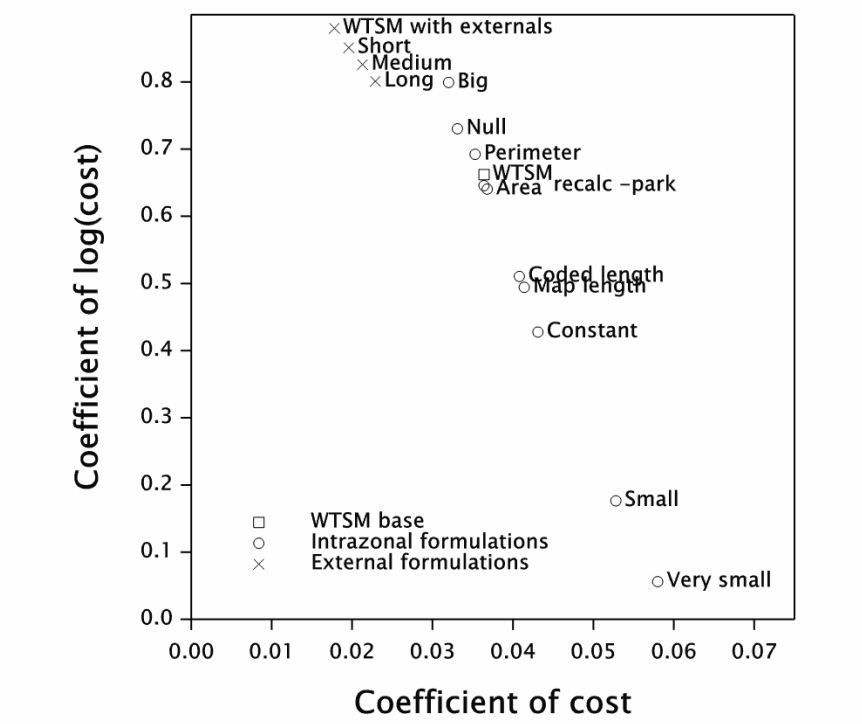
Intrazonal cost formulation	Exponential		Tanner						Power	
	Cost		Cost		Ratio		Log(cost)		Log(cost)	
	Coef	t	Coef	t	Coef	t	Coef	t	Coef	t
Zero	0.06329	28.93								
Very small	0.06329	28.93	0.0581	23.94	1.0	4.04	0.0561	4.46	0.320	39.92
Small	0.06331	28.92	0.0528	18.94	3.3	4.60	0.1764	5.61	0.738	43.94
Constant	0.06345	28.84	0.0431	12.89	9.9	4.89	0.4276	7.32	1.173	43.61
Big	0.06375	28.54	0.0320	8.25	25.0	4.38	0.7992	8.72	1.560	41.05
Perimeter	0.06336	28.46	0.0353	9.85	19.6	4.82	0.6922	8.77	1.462	42.23
Area	0.06377	28.54	0.0368	10.45	17.4	4.95	0.6404	8.71	1.397	43.15
Coded length	0.06371	28.74	0.0408	12.12	12.5	5.10	0.5104	8.13	1.253	44.55
Map length	0.06377	28.73	0.0414	12.53	11.9	5.21	0.4943	8.22	1.225	44.83

The coefficients of the Exponential function are remarkably stable. The only difference of note is for the null formulation, demonstrating that intrazonals do have some effect on the model's fit.

The coefficients of the Power function are much less consistent. The null formulation is at one end of the range, suggesting that none of the formulations of intrazonal costs allows a single Power model to fit both intrazonals and interzonals well. At the other end of the range is the very small formulation, suggesting that it is not a good or sound fix for the Power function's inability to handle zero costs.

The coefficients of the Tanner function also vary considerably, but are closely correlated with each other, as is shown in figure 4.47.

Figure 4.47 Coefficients fitted to the Tanner function



The outlying points at the bottom right represent the very small and small formulations, again suggesting that these are not satisfactory approximations to zero costs.

Although the ratio between the coefficients varies, its t value shows it is always strongly positive. Thus no formulation is fitted by a Tanner function with a turning point or hump.

4.11.3 Measures of fit

Table 4.35 shows the deviance over all movements, both intrazonal and interzonal. Intrazonals are excluded from the null formulation, so its deviance is not comparable.

Table 4.35 Residual deviances for alternative intrazonal cost formulations

Intrazonal cost formulation	Exponential	Tanner	Power
WTSM	4250	4179	4355
– less parking charges	4250		
– and recalculated intrazonal costs	4249	4183	4353
Null (excludes intrazonals)	*4110*	*4063*	*4168*
Zero	4240		
Very small	4240	4221	5371
Small	4240	4211	4843
Constant	4243	4193	4458
Big	4259	4192	4296
Perimeter	4277	4211	4370
Area	4259	4194	4376
Coded length of centroid connectors	4247	4190	4442
Map length of centroid connectors	4246	4188	4463
Range	37	42	1075

Again, the Exponential function is relatively insensitive, the Power function is very sensitive, with the Tanner closer to the Exponential's stability.

Small constants for intrazonal cost fit slightly better than the WTSM formulation with the Exponential, but much worse with the Power, where the big formulation fits best. Otherwise, the WTSM formulations generally fit as well as any.

These deviances show the Exponential always fitting better than the Power function, contrary to the t statistics in table 4.34.

4.11.4 Fitted trips and travel

Table 4.36 shows the fitted trips and travel costs for interzonal movements. These are the movements that appear on the links in the model that represent the real network. The observed number of trips is always replicated when intrazonal movements are included, so excluding them shows the division of trips between interzonal and intrazonal movements. The amount of intrazonal travel depends on the formulation of intrazonal costs. Travel is in generalised minutes and in the original WTSM formulation includes parking charges.

Table 4.36 Interzonal trips and travel

Intrazonal cost formulation	Exponential		Tanner		Power	
	Trips	Travel	Trips	Travel	Trips	Travel
WTSM (incl. parking charges)	172,337	4,415,154	166,568	4,401,023	159,719	5,392,062
– less parking	172,337	4,123,906				
– recalculated	172,348	4,123,647	166,902	4,110,303	160,319	5,068,543
Null	167,285	4,109,409	167,285	4,109,409	167,287	4,753,276
Zero	171,069	4,109,353				
Very small	171,069	4,109,353	166,499	4,109,351	157,369	8,626,247
Small	171,112	4,109,736	165,554	4,109,178	152,460	6,741,455
Constant	171,504	4,113,574	165,208	4,107,277	155,322	5,387,449
Big	173,186	4,138,867	170,492	4,125,398	169,077	4,762,364
Perimeter	173,948	4,178,636	169,720	4,165,446	164,494	5,113,963
Area	173,148	4,145,728	167,975	4,132,575	161,200	5,181,644
Coded length	172,138	4,119,685	166,421	4,110,357	157,536	5,453,834
Map length	172,181	4,119,373	166,368	4,110,341	157,106	5,567,479

In the observed matrix, there were 167,283 trips and 4,109,333 generalised minutes of travel for interzonal movements.

The Exponential deterrence function generally overestimates these trips, and to a much lesser extent the travel. The Tanner function does not show the same bias and gives closer matches. The Power function underestimates trips to a greater extent than the Exponential overestimates them, and shows a serious overestimation of travel that is sensitive to the formulation of intrazonal costs.

Since the null formulation is fitted to interzonal movements alone it replicates both trips and travel, except for travel under the Power function. This is because the Power function fits only the logarithm of cost.

4.11.5 Fitted formulations for intrazonal costs

The formulations above use fixed coefficients to calculate intrazonal travel costs, eg √area/2 or perimeter/8. GLMs can be specified to estimate the coefficients that give the best fit to the model by fitting a separate cost coefficient or L factor to intrazonal movements. A constant term can also be fitted by a K factor. Either a K or L factor gives the same model for all cases of constant intrazonal cost, ie the zero (K only), very small, small, constant or big formulations.

Table 4.37 shows the fit of models with both cost coefficients and constants (K and L factors) fitted separately to intrazonal movements. The fit is expressed as residual deviance, and its change from table 4.35.

Table 4.37 Fit with separate intrazonal factors

Intrazonal cost formulation	Exponential		Tanner	
	Residual	Change	Residual	Change
WTSM, recalculated less parking	4226	23	4181	2
Constant	4227	16	4183	11
Perimeter	4226	51	4186	25
Area	4226	34	4183	11

Intrazonal cost formulation	Exponential		Tanner	
	Residual	Change	Residual	Change
Coded length of centroid connector	4225	21	4184	6
Map length of centroid connector	4224	22	4183	5

With both K and L factors included, intrazonal and interzonal movements are to a large extent fitted separately; they are linked only through trip end balancing factors and a common dispersion. The fit of interzonal movements becomes largely independent of the formulation of intrazonal costs, so the overall residual deviances become quite similar.

With the Exponential deterrence function, extra factors for intrazonals are quite significant but often take perverse values – for example, negative coefficients of intrazonal cost. On examination, the factors are accommodating the Exponential's general underestimation of short trips in the manner shown in figure 4.28. Tanner functions address this issue with a curved form; their basic fit (table 4.35) is always better than Exponential functions with extra intrazonal factors.

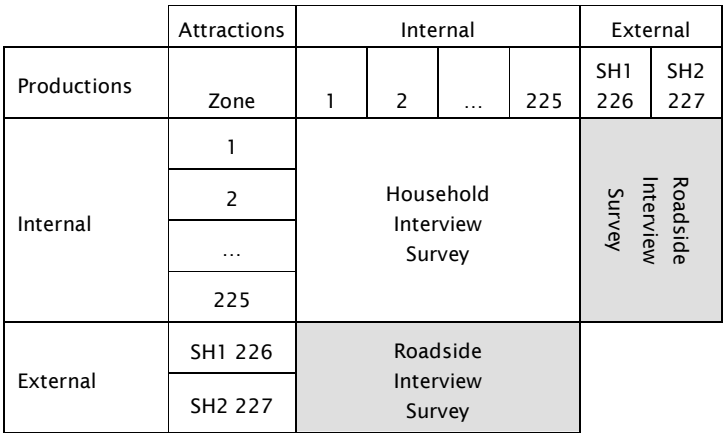
Because the Tanner and Power deterrence functions involve transformations of costs, simple application of K and L factors give only approximate coefficients for intrazonal costs. Such simple application to the Tanner function produces relatively little improvement in the WTSM formulation (recalculated without parking). Other formulations improve more, but only to the level of the WTSM formulation.

Exact coefficients for the best-fitting intrazonal cost formulations require non-linear specifications with Tanner or Power deterrence functions. Such specifications may be fitted by extensions of GLMs, but are probably not justified for the crude measures used here to investigate the sensitivity of deterrence. The plain WTSM formulation is about as good as any tried here, if used with a Tanner deterrence function.

4.12 Sensitivity to external zone costs

The rest of this study analyses the distribution of trips internal to the study area, observed in the household interview survey (HIS). The WTSM distribution model also includes external trips derived from roadside interviews. As discussed in section 1.5.4, these were omitted to focus on a single consistent data source, and because of apparent inconsistencies that may have arisen from interview site locations or the coding of external connectors. This section briefly examines alternative costings for the external connectors and the sensitivity of the fit of different deterrence functions to them.

Figure 4.48 WTSM trip matrix observed from household and roadside surveys



In this section, and the study generally, external trips refer to trips with one end outside the study area and one inside – the shaded areas of figure 4.48. The size and topography of the study area and its sparse external links make external-to-external commuter trips, with both trip ends outside the study area, very unlikely to pass through the study area. There are effectively just two external zones for commuting by car, zone 226 on State Highway 1 (SH1) and zone 227 on SH2. A third external zone, 228, represents the South Island, linked by the Cook Strait ferries.

4.12.1 External cost formulations

In the WTSM model network, the external zones are attached by centroid connectors 5km long, coded with a speed of 40km/h. Actual centres of population and employment lie further beyond the study area boundary. Table 4.38 shows plausible centres for commuting in and out of the study area. These are taken to define short, medium and long formulations of alternative external costs.

Table 4.38 External costs

External cost formulation	Zone 226 on SH1				Zone 227 on SH2			
	Location	Distance km	Time min	Cost gen min	Location	Distance km	Time min	Cost gen min
WTSM	nominal	5	7.5	12.13	nominal	5	7.5	12.13
Short	Levin	15.4	12	26.27	Eketahuna	27.3	18	43.30
Medium	Shannon or Foxton	32	24	53.66	Pahiatua	52.4	36	84.57
Long	Palmerston North	64.1	52	111.41	Palmerston North	89.3	73	155.77

Distances and times are taken from driving directions in Google mapping, measured from the last nodes on the WTSM highway network. These are nodes 1651 (2695147m east 6050123m north) and 1727 (taken at 2732370m east 6035808m north, old New Zealand grid) on SH1 and SH2 respectively. Costs are then calculated according to the standard WTSM formulation (appendix A). There is roughly a doubling of costs between successive formulations. The cost of travel to Palmerston North is much more than is typical of commuter trips within the study area, shown in figure 4.1.

The alternative costs were added to external movements in cost matrices and the nominal WTSM values subtracted. The cost matrices included parking charges in the CBD, which were suppressed for the sensitivity to intrazonal costs considered in the previous section.

As in the WTSM calibration, the 12 external trips reported in the household survey were simply omitted and no attempt has been made to reconcile them with roadside interview data. Weighting for the external trips was set at 1/4.02 from table 3.11, allowing for uneven sampling. Weighting of the internal trips remained at 1/157.9, reflecting the much lower sampling rate of independent workplaces in the HIS. Applying these weights to the expanded samples of trips in table 4.39 shows that external trips make up almost half of the effective sample in the combined dataset for calibration, but only 2% of trips in the study area.

Table 4.39 Sample of internal and external commuter car trips

Scope	Observed	Expanded	Weight	Effective sample
Internal	3045	183,215.7	1/157.9	1160.3
External	1190	3901.3	1/4.02	970.5
Combined	4235	187,117.0		2130.8

Inclusion of the external movements increases the number of matrix cells in the dataset because the roadside surveys intercepted trips to or from internal zones for which no trips were sampled in the HIS. This reduces the number of empty zones (which are omitted from analysis) in the combined sample, as shown in table 4.40. The matrix is also extended by the two external zones.

Table 4.40 Non-empty zones and cells

Scope	Non-empty zones		Matrix cells in dataset
	Production	Attraction	
Internal	162	194	31,428
External	114	113	454
Combined	183	202	36,962

For the internal dataset, the number of cells is simply the product of the non-empty production and attraction zones.

The non-empty zones counted for the external movements are internal zones. Each contributes two cells to the dataset, corresponding to the two external zones or the shaded area of figure 4.48. Preliminary fitting of distributions to these external trips alone suggested a surprisingly powerful dataset given its limited scope, if not its effective sample size.

The four movements within and between the two external zones were suppressed by omitting them from the dataset, as in a partial matrix. This is the difference between the product of non-empty zones and the number of cells in the combined dataset, in the bottom row of table 4.40; it is the missing bottom right corner of figure 4.48.

The inclusion of the external trip sample increases the number of matrix cells analysed by 18%.

4.12.2 Fitted models

Table 4.41 shows the coefficients for Exponential, Tanner and Power deterrence functions calibrated with the alternative external costs. The first line is a null formulation that omits all external movements and is the fit to internal movements calibrated previously.

Table 4.41 Fitted coefficients for alternative external costs

External cost formulation	Exponential		Tanner						Power	
	Cost		Cost		Ratio		Log(cost)		Log(cost)	
	Coef	t	Coef	t	Coef	t	Coef	t	Coef	t
Null	0.0638	28.68	0.0364	10.44	18.2	5.04	0.662	9.01	1.416	43.29
WTSM	0.0336	36.83	0.0178	16.23	49.5	5.17	0.880	19.42	1.478	54.38
Short	0.0336	36.83	0.0196	19.36	43.3	4.03	0.851	19.59	1.497	53.54
Medium	0.0336	36.83	0.0213	22.31	38.8	3.28	0.826	19.76	1.487	52.31
Long	0.0336	36.83	0.0229	25.27	34.9	2.73	0.801	19.91	1.445	50.78

There is a very marked change in the Exponential coefficient when external movements are also considered, reflecting a discrepancy between internal and external trip distributions. The decrease in the coefficient with the inclusion of longer external trips is consistent with Daly's cost damping (2010) and with the contrast between urban and regional models noted by Bly et al (2001, section 8.3). Once external

movements are included, the Exponential model is not affected by common changes to all access costs to a zone, which are absorbed in its trip end balancing factors.

The coefficient of the Power function is much less sensitive to the inclusion of external trips and quite insensitive to their formulation.

The Tanner function broadly follows the patterns of its Exponential and Power components, with the coefficient of cost more sensitive to external costs than that of log(cost). Figure 4.47 shows the usual trading-off between the coefficients with less spread if not along exactly the same line as the intrazonal formulations shown in the same plot.

4.12.3 Measures of fit

Residual deviances are shown in table 4.42.

Table 4.42 Residual deviances for alternative external costs

External cost formulation	Exponential	Tanner	Power
WTSM	5302.7	4984.9	5343.2
Short	5302.7	4977.4	5529.1
Medium	5302.7	4969.3	5750.7
Long	5302.7	4961.6	6023.5

The insensitivity of the Exponential deterrence function to access cost is again apparent. The Power function is sensitive, but its fit deteriorates as external costs increase through their likely range. The Tanner is less sensitive by at least an order and its fit improves as external costs increase from the nominal WTSM values. However, there is no minimum in the likely range for commuting across the study area boundary.

Over the whole of the range tested, the Tanner fits better than the Exponential, and the Power worse, as has been found for the internal dataset alone.

4.12.4 Fit of separate coefficients for internal and external movements

Preliminary fitting with an Exponential deterrence function found a marked difference between the distributions of internal and external trips. This is apparent from the difference in coefficients between the null formulation and others at the top of table 4.41. To test whether this affects other functions over the range of external costs, separate coefficients were fitted for internal and external movements. For the Tanner, separate coefficients were fitted for the cost term but not the log(cost) term, so there is only one more degree of freedom in all the deterrence functions. The separate coefficients are effectively L factors, or parameters in WTSM terminology. Separate constants, or K factors, are inherent in the trip end balancing factors.

The residual deviances and their change from those with a common coefficient (table 4.42) are shown in table 4.43.

Table 4.43 Residual deviances with separate coefficients for internal and external movements

External cost formulation	Exponential		Tanner		Power	
	Residual	Change	Residual	Change	Residual	Change
WTSM	5008.2	294.5	4944.7	40.2	5161.8	181.4
Short	5008.2	294.5	4947.9	29.5	5170.8	358.2
Medium	5008.2	294.5	4947.8	21.5	5171.0	579.7
Long	5008.2	294.5	4946.5	15.1	5161.7	861.8

With differences between the internal and external distributions thus simply accommodated, the residual deviances are hardly sensitive to the external costs; the Exponential is again insensitive. The Power and Tanner both show a maximum in the range of external costs, where a minimum would be expected if the formulation of external costs were critically affecting the fit.

The improvement over a common distribution is generally most marked for the Power function and least marked for the Tanner. Even for the Tanner, the improvement is significant. Further separation of the internal and external distributions may be complicated by the appearance of empty zones in the individual datasets and different levels of sparsity according to the sample rate. Differences may have arisen from the location and timing of the roadside interviews, determined by practicalities of safety and the weather.

The Tanner remains the best-fitting function over the range of external costs and the Power the worst. Perhaps most importantly, a Tanner function with a common distribution (table 4.42) always fits significantly better than an Exponential or Power function with separate coefficients. It accounts simply for all of the discrepancy between internal and external distributions fitted here in the Exponential models.

It is possible that separate external zones at the short, medium and long distances from the study area could still improve the fit. However, this does not seem promising from the results so far and it would require re-coding of the roadside interview data. Other deterrence functions may resolve some of the discrepancies that remain with the Tanner.

Depth in external zoning may be more important in smaller study areas, with larger proportions of external and through flows. External trips are often treated separately in practice, if only because the volume and assignment of trips generated in them are not fully modelled.

4.13 Summary

This analysis is of one study area (Wellington), one purpose (HBW), one mode (car), generally from one survey (household, internal trips only). The large study area, with its wide range of trip costs, may allow small differences in deterrence functions to appear statistically significant; ie the dataset may have greater power than one for a smaller study area with the same sample size.

Exponential

The Exponential has been taken as the base function, appearing as a straight line when log(deterrence) is plotted against cost. There is evidence that concave curvature in the line fits better, with relatively more trips at high and low costs.

Power and Tanner

Although the Power function produces such curvature, it fits markedly worse than the Exponential.

Combining the Power and Exponential functions in the Tanner function does improve considerably on the Exponential, with a concave curve fitted.

Empirical functions

The traditional step form does not produce an efficient fit. A good fit requires many degrees of freedom and most of these seem to be needed to approximate the steps to a continuous function. The form does not automatically reproduce mean trip costs: it requires subdivision of the higher-cost ranges, in which there are few trips, to give a good approximation. The form is computationally convenient if software for fitting other forms is not available.

GLMs allow several other forms of empirical functions, segmented by cost. Fitting a common slope to all the segments of the traditional step form, in effect adding an Exponential component, always improves the fit significantly and reproduces mean trip costs.

Separate slopes for each segment can be forced to join, eliminating steps between straight lines, using 'broken-stick' functions. Since this is another derivative of the Exponential, it must improve its fit, but the improvements for a few breaks in slope are very significant. A single break in slope does not fit as well as the Tanner function, which has the same degrees of freedom to fit a continuously curving line.

Larger numbers of changes in the slope, or allowing steps at the break in slopes, do not produce such significant improvements in fit.

The main departures from the straight line of the Exponential are towards a concave curve.

Geographic segmentation

This form of segmentation, adopted in the WTSM, shows the same signs of decreasing slopes at higher costs. It does not fit as well as segmenting simply by cost. Separate coefficients (L factors for slopes) produce a better fit than separate constants (K factors for steps). However, even with both factors fitted separately over five geographic segments (8df), the fit is not quite as good as the Tanner function (1df).

Most of the improvement in fit from geographic segmentation arises from allowing curvilinearity in the basic Exponential cost function.

The WTSM distribution model incorporates simultaneous mode split which may benefit more directly from geographic segmentation, eg in representing rail commuting to the CBD.

Splines

Splines are a sophisticated form of empirical function that can be fitted in GLMs. They are similar to the multi-segment slopes, but the break points are determined in the fitting process and the change of slope is blended into a continuous curve.

Splines could be fitted with many degrees of freedom; models with up to 50df were tried. These showed turning points in the function implying more travel at greater cost. The improvements in fit beyond the first few orders were not significant.

Polynomials

Polynomials are a more traditional approach to investigating curvature in linear regression. Like splines, they could be fitted to higher orders where improvements were not significant. The pattern was more irregular than for splines, possibly due to alternation in the form of odd and even polynomials. Turning points appeared from low orders and computational instability appeared before breakdown between 13df and 21df.

Unlike splines, polynomials continue to curve beyond the limits of fitted data, making extrapolation riskier.

Both splines and polynomials were tried with the logarithm of cost to reduce the range of higher costs with sparse data. This did tend to stabilise the function at high costs, but allowed more fluctuation at low costs. The Tanner function, combining cost and log(cost), was taken as a starting point for some of the analyses.

Non-linear functions

Non-linear functions that fall outwith standard GLMs can be fitted by an extension to the GLM algorithms in Genstat. Several forms such as the Box-Cox and log-normal fit better than the linear forms of the Exponential and Tanner. Some special cases (eg fixed offsets) can be reformulated as linear functions. Only the Box-Tukey function could be fitted with two non-linear coefficients, and the linear coefficient of

the non-linear function could not be suppressed to fit some functions. The fitting and interpretation of non-linear models is more difficult than of standard GLMs.

Advances on the Tanner function

Compared with the Tanner function, splines, polynomials and non-linear functions tend again to show concave curvature, with more travel at higher cost, over and above the similar changes introduced by the Tanner over the Exponential.

There was little indication of change in the function for the middle range of costs, where most trips were observed, or for low costs which can affect the fitting of intrazonals.

Measures of statistical sgnificance

From empirical observation of modelling this dataset:

Mean residual deviances from almost all models were less than the expections calculated by the elaboration of Poisson probabilities for the fitted means. Because of this dependency on fitted means, the expected residuals vary by model, and tend to decrease with the observed mean residual. Mean residual deviances appear to reflect as much on the adequacy of the error model, represented by the weighting scale, as they do on the fit of the systematic model.

The change in residual deviance when adding a term appears to be a useful statistic. When overfitting models, the changes are close to one per degree of freedom, as expected for a χ^2 distribution when random terms are added.

The t statistic (mean/standard error) is often similar to the square root of the change in deviance for marginally significant changes ($t \approx 2$ or $\chi^2_1 \approx 4$), but can differ for larger changes. The t statistic appears relatively small for the first natural cost term in a model (eg an Exponential model), and can differ greatly from the change in deviance in non-linear models. The change in deviance is preferred for choice of models.

The Pearson chi-squared test, or Poisson Index of Dispersion, is a thoroughly unreliable statistic for this sparse data.

Sample sizes

The prime effects of cost deterrence are so strong that they could be detectable in small samples, of a hundred or fewer households or trips. Even the improvement from the Exponential to the Tanner (adding log(cost)) can be detected from a modest sample of hundreds, so it might be apparent in a survey on the scale of the WTSM (2538 households) in other purpose distributions if it occurs to the same extent.

Practical measures of fit

Models with the various deterrence functions have also been compared in terms of screenline crossings and the effects of road improvement schemes.

Differences in fit vary more between screenline and scheme locations than between deterrence functions. There is no clear benefit in these terms from more sophisticated deterrence functions than the Exponential.

This suggests a geospatial error component which cannot be addressed by costs.

Generalised costs

The analyses above have been based on the generalised cost formulation adopted in the WTSM for trip distribution. Its components – time and distance, AM and IP – from the assignment stage have been examined. With an Exponential deterrence function, the distance component of generalised cost does not improve the fit of the distribution model significantly and the IP times provide an adequate model. These

differences are relatively small. Simpler measures of separation such as crow-fly distances and intervening opportunities fit significantly less well, but still capture much of the distribution effect.

Intrazonal costs

The Tanner model fits interzonal trips and travel cost, which appear on the 'real' network, better than the Exponential, but is slightly more sensitive to the formulation of intrazonal costs. The Power function is worse than the others in both respects.

The WTSM formulation for intrazonal costs (half minimum interzonal, capped) gives as good a fit as any tried.

External costs

The Tanner function can account for all of the discrepancies between internal and external trip distributions found in Exponential models. There remain some unresolved discrepancies between the two datasets, derived from household and roadside interviews respectively.

Power functions are very sensitive to the specification of external costs, the Tanner function much less so, and the Exponential not at all.

No evidence has been found that realistic external costs improve the fit of any of these models, but the size and topology of the WTSM study area render external movements relatively unimportant compared with most study areas.

5 Disaggregate modelling

Analysis so far has been at the zonal level of aggregation, as is conventional in trip distribution modelling. Much recent development in choice modelling has been at the disaggregate level of individual trips. This has become a practical approach to modal and more recently time-of-travel choice by multinomial or nested logit models. However, these present relatively few choices, and the many choices between attraction zones have presented computational problems for calibrating trip distribution by disggregate methods.

There are two distinct intermediate levels of aggregation between zone and trips, namely households and persons. These are apparent in the HIS data structure, which has related tables for households, persons and trips. Variables held at a level, such as occupation for a person, are naturally modelled at that level. The effect of disaggregation is explored by fitting extra variables that are related to these levels and whose full contrasts do not appear after aggregation to zones. It demonstrates links between the conventional aggregate form of the gravity model and the disaggregate form of destination choice modelling, with intermediate levels of aggregation.

Conceptually, the production zones are progressively divided and redefined from geographic areas into individual households, individual persons and then individual trips. In the final matrix, disaggregated to trips, there is one row representing every trip. All the cells in that row are zero except for the zone to which the trip is attracted. Computationally, each such row is generated from one record in the trip database and rows are aggregated as required to persons, households or the original geographic zones. All forms are fitted with a common algorithm and with a full set of 194 attraction choices, without sampling.

5.1 Variable selection and preparation

Variables were chosen for:

- ready availability and consistent coding in the survey database
- being plausible influences on trip distribution
- occurring at different levels of aggregation – household, person and trip.

The variables suggest a wide variety of interactions that could influence travel demand. These might be represented by activity or tour modelling, and influence trip generation and mode choice as much as trip distribution. It was not the intention to develop sophisticated models of these variables' effects, but to use them as case studies in disaggregation to test the process.

Each variable is described by a single value, either a continuous variable (eg age in years or income in dollars), or as a dichotomy (eg male/female or white/blue collar). This simplifies tabulation and appraisal, but particularly aggregation: dichotomies can be represented by dummy variables that can be simply averaged. (Genstat is adept at handling factors with multiple levels, but not at averaging them; sets of averaged dummies could be handled by pointers, but not with the same dexterity.) Dichotomies were chosen for a balanced 50:50 split in the dataset where possible and coded 1 for the case thought likely to travel further.

When aggregating from individual trips to persons, households and zones, averaging of the variables is always simply by trips. Other averaging schemes are possible; equal weight per person rather than trip would be consistent with the error model. A stepwise scheme could give equal weights to trips within person, person within household and household with zone. This raises the issue of non-mobile persons and households, which are automatically excluded when averaging by the trips made. Other averaging schemes would not necessarily give the same consistency of results found here.

Cost is not affected by the averaging method up to zones, because it is taken as the same within any zone. Aggregation of different costs across zones is considered in section 7.6.

Sources of variables are defined in terms of the variable names used in the WTSM household interview survey. These names are prefixed by h, p, or t for variables at the household, person or trip levels of disaggregation and are held in the corresponding tables in a relational database, eg:

hNumVisitors	Household	Number of visitors
pOccupation	Person	Classification of occupation
tOriginType	Trip	Type of land-use activity at origin

5.1.1 Zone level

Cost (or separation) is the key variable in trip distribution. In transport modelling it is usually calculated at the zonal level. Daly and Ortuzar (1990) considered further refinement. Variations in the definition of cost have already been examined in section 4.10; these included time and distance, and AM and IP conditions. The costs adopted here are the standard WTSM formulation for private mode work trips in the demand model, but with CBD parking charges removed in anticipation of modal and spatial modelling, and with intrazonal costs recalculated as described in section 4.11.1.

5.1.2 Household level

5.1.2.1 Household segment – car availability

This variable represents the number of vehicles available to the household, compared with the number of adults in it. It is used to segment the WTSM distribution and mode split model for HBW.

In the WTSM it has three levels. For this analysis, the captive segment with no vehicles was combined with the competition segment, with fewer vehicles than adults. The dummy variable represents the choice segment, having at least one vehicle per adult.

5.1.2.2 Single worker

This dummy variable indicates households with only one worker. Workers are counted whether they make trips or not, so multi-worker households may be represented by only one worker in the dataset of trips.

From pHrsEmployeeType not null.

5.1.2.3 No children

This dummy represents the absence of children.

From hNumResidents+hNumVisitors–Adults = 0; Adults from pBirth<=1984, as used for household segmentation.

5.1.2.4 Income

Household income is the sum of incomes of all persons in the household. See section 5.1.3.5.

5.1.2.5 Also considered

Other measures of household structure, possibly from category analysis.

5.1.3 Person level

5.1.3.1 Gender

The dummy variable represents males.

5.1.3.2 Age

Age is treated as a continuous variable.

It is derived from pBirth because pAge has missing values.

5.1.3.3 Occupation

The dummy variable represents technical, professional and other occupations with codes less than 40 in the table below. This gives a roughly equal division of trips, thanks to the presence of so many corporate managers. Inclusion of clerical and sales would be a better definition of white-collar occupations, but this would give a 22:78 split and probably a higher correlation with the industry variable.

Table 5.1 Occupations

Sector	Occupation	Code	Persons	HBW car trips		
				No.	%	cumulative
Not a worker		~	3396	91	2.99	3.0
Armed forces	Armed forces	1	1	0	0.00	3.0
Administrators & managers	Legislators & administrators	11	17	17	0.56	3.5
	Corporate managers	12	469	475	15.60	19.1
Professionals	Physical, mathematical & engineering	21	171	130	4.27	23.4
	Life science & health professionals	22	131	117	3.84	27.3
	Teaching professionals	23	192	198	6.50	33.8
	Other professionals	24	288	184	6.04	39.8
Technicians & associate professionals	Physical & engineering associate	31	132	93	3.05	42.9
	Life science & health associate	32	31	34	1.12	44.0
	Other associate professionals	33	326	263	8.64	52.6
Clerks	Office clerks	41	295	230	7.55	60.2
	Customer service clerks	42	114	108	3.55	63.7
Service & sales workers	Personal & protective services	51	366	269	8.83	72.5
	Salespersons, demonstrators	52	210	152	4.99	77.5
Agriculture	Market oriented agricultural & fishery	61	118	72	2.36	79.9
Trades workers	Building trades workers	71	172	140	4.60	84.5
	Metal & machinery operators	72	70	95	3.12	87.6
	Precision trades workers	73	33	37	1.22	88.8
	Other craft & related trades workers	74	40	27	0.89	89.7
Plant & machine operators & assemblers	Industrial plant operators	81	20	16	0.53	90.2
	Stationary machine operators & assemblers	82	61	73	2.40	92.6
	Drivers & mobile machinery operators	83	75	70	2.30	94.9
	Building & related workers	84	10	6	0.20	95.1
	Other	89	1	1	0.03	95.2
Elementary	Labourers & related elementary services	91	195	133	4.37	99.5
Unknown	Not enough information given	97	14	10	0.33	99.9
	Refused to answer	99	5	4	0.13	100
Total			**6953**	**3045**	**100.00**	

From pOccupation<40. Nulls and missing values from non-workers (lift-givers?) are treated as non-professionals.

5.1.3.4 Industry

The WTSM fitted different attraction rates for service industries (TN 16.3) so these have been adopted as the variable for this analysis.

Table 5.2 Industries

Sector	Industry	Code	Persons	HBW car trips		
				No.	%	cumulative
Not a worker		~	3396	91	2.99	3.0
Agriculture, hunting, forestry & fishing	Agriculture & hunting	11	111	65	2.13	5.1
	Forestry & logging	12	5	5	0.16	5.3
	Fishing	13	5	0	0.00	5.3
Mining & quarrying	Crude petroleum & natural gas	22	2	0	0.00	5.3
	Metal ore mining	23	1	2	0.07	5.4
	Other mining & quarrying	29	1	1	0.03	5.4
Manufacturing	Food, beverage, tobacco	31	52	54	1.77	7.2
	Textile, apparel & leathergoods	32	28	20	0.66	7.8
	Wood processing & wood products	33	23	23	0.76	8.6
	Paper products; printing & publishing	34	61	54	1.77	10.3
	Chemical products	35	35	42	1.38	11.7
	Mineral products	36	10	9	0.30	12.0
	Basic metal industries	37	8	6	0.20	12.2
	Fabricated metal products, machinery	38	112	128	4.20	16.4
	Other manufacturing industries	39	9	5	0.16	16.6
Electricity, gas & water	Electricity, gas & steam	41	20	10	0.33	16.9
Construction	Construction of buildings	51	134	124	4.07	21.0
	Construction other than buildings	52	15	20	0.66	21.6
	Ancillary construction services	53	95	63	2.07	23.7
Wholesale & retail trade and restaurants & hotels	Wholesale trade	61	91	69	2.27	26.0
	Retail trade	62	421	374	12.28	38.3
	Restaurants & hotels	63	184	126	4.14	42.4
Transport, storage & communication	Transport & storage	71	124	108	3.55	45.9
	Communication	72	84	73	2.40	48.3
Business & financial services	Financing	81	128	99	3.25	51.6
	Insurance	82	48	36	1.18	52.8
	Real estate & business services	83	453	338	11.10	63.9
Community, social & personal services	Public administration & defence	91	324	260	8.54	72.4
	Sanitary & cleaning services	92	31	22	0.72	73.1
	Social & related community services	93	685	601	19.74	92.9
	Recreational & cultural services	94	140	111	3.65	96.5
	Personal & household services	95	111	101	3.32	99.8
Unknown	Refused to answer	99	6	5	0.16	100
Total			**6953**	**3045**	**100.00**	

From pIndustry>60. Non-workers treated as not service industry.

5.1.3.5 Income

Income is converted to a continuous variable, in NZ$/year, as the midpoints of the intervals in which pAdjIncome is coded. The bottom band (loss) is taken as -$5000pa, and the top band as $120,000pa. Refusals and other missing data were infilled during WTSM development (TN9.1 'Survey processing', section 3.5.2).

5.1.3.6 Also considered

The 'primary worker' in a household was identified in preparing land-use formulations (appendix B).

5.1.4 Trip level

It was hard to select sensible variables for individual trips without considering their part in a tour, but this would have led to extensive further analysis. Mode is a key characteristic at trip level which was not being considered at this point; again, it acts largely at a tour level and is influenced by household car availability and activities.

5.1.4.1 Direction of travel

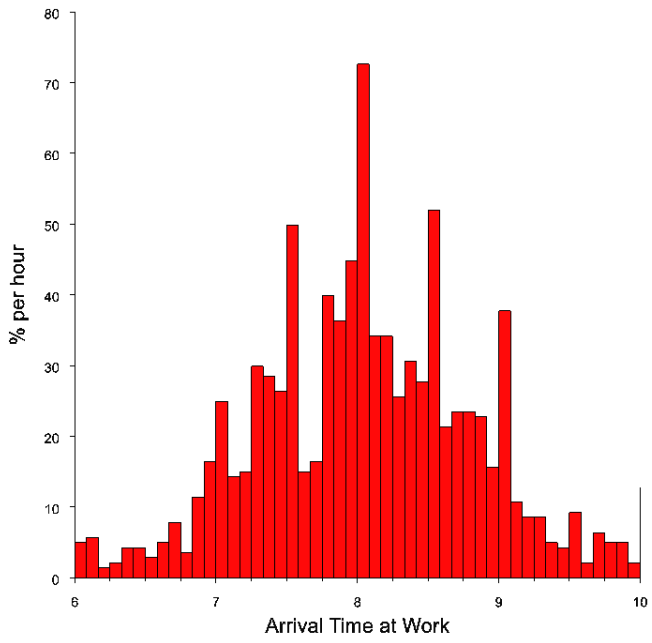
The dummy variable represents outward travel from home to work. Outward and return trips are naturally balanced, to and from the same place, but imbalance may appear from intermediate stops on one leg or the other.

From tOriginType = 10 (home)

5.1.4.2 Peak travel

The dummy variable represents travel in the peak periods, defined as arrivals at work between 7am and 9am and departures between 4pm and 6pm, inclusive of all those times. More extensive disaggregate modelling of time choice could define shorter time segments through the peak and its shoulders, and consider the time of day and duration of stay required at the workplace.

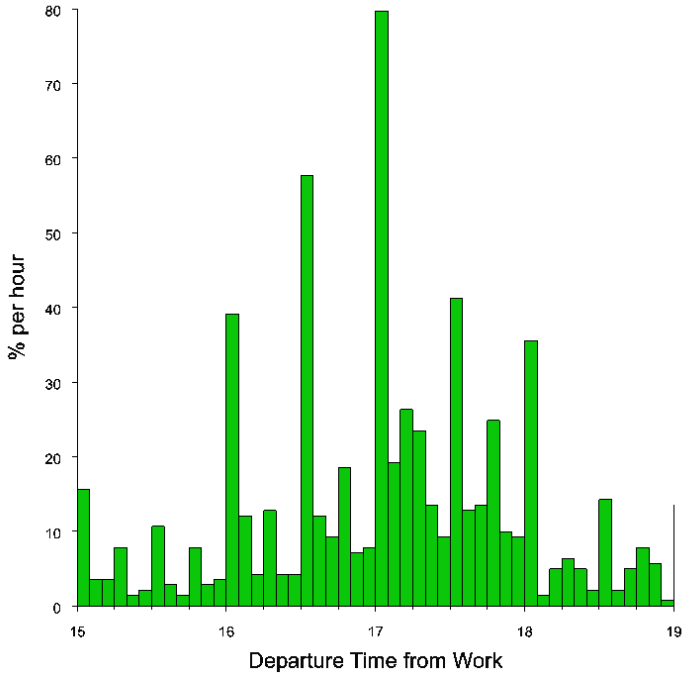
Figure 5.1 Arrival time at work



From tDestTime

Times were recorded to the minute, but tended to be given to the round five minutes or major fractions of the hour, particularly for departure. In these histograms, horizontal bands are for five minutes, including their lower limit and excluding their upper one, eg 7am to 7.04am inclusive. The vertical axes show hourly rates of arrivals or departures, as percentages of the daily total.

Figure 5.2 Departure time from work



From tOriginTime

5.1.4.3 Car sharing

The dummy variable represents sharing the car.

From tHhldNum+tNonHhldNum>1; missing taken as not sharing.

5.1.4.4 Also considered

Parking types and fees are coded only to the destination end of trips, so they usually appear only for the outward trip. Even there, only 10% or so involve payments. The WTSM includes a parking charge in costs to attraction zones in the CBD only. Use of company or other non-household cars is also a minority.

5.1.5 Summary of variables

Table 5.3 summarises these variables. The first column shows the proportions of all HBW car trips, or averages of continuous variates weighted by these trips.

Table 5.3 Means and contrasts of variables

Variable	Global mean	Standard deviation between averages			
		Zone	Household	Person	Trip
Number of cases		162	1224	1740	3045
Trips per case		18.80	2.49	1.75	1
Car available	0.60	0.24	**0.49**	0.49	0.49
One worker	0.22	0.16	**0.41**	0.41	0.41
No children	0.58	0.20	**0.49**	0.49	0.49
Household income	$81,976	$22,731	**$43,364**	$43,364	$43,364
Age	41.0yrs	5.3yrs	11.7yrs	**13.2yrs**	13.2yrs
Male gender	0.56	0.13	0.37	**0.50**	0.50
Professional occupation	0.50	0.21	0.44	**0.50**	0.50
Service industry	0.76	0.17	0.37	**0.43**	0.43
Personal income	$39,699	$11,481	$22,283	**$25,548**	$25,548
Outward	0.55	0.06	0.20	0.24	**0.50**
Peak	0.58	0.15	0.36	0.41	**0.49**
Share	0.27	0.16	0.38	0.42	**0.45**
Random sample	50:50	0.12	0.32	0.38	**0.50**
	25:75	0.10	0.27	0.33	**0.43**

Bold – level of disaggregation at which the variable occurs

The remaining columns show the differences between such averages aggregated at different levels – zone, household, person or trip. The number of each of these cases is shown in the first row and the average number of trips per case in the second. As the variables are averaged over more trips for each case, towards the left of the table, the difference between them in the main body of the table grows less. Any effects of the variables have to be found from these contrasts between cases, which tend to be lost in aggregation.

The main body of the table is divided into three sets of rows, according to the level of aggregation at which variables occur – household, person and trip. The corresponding column is shown in bold. At greater disaggregation, to the right of the bold figures, the variables are simply replicated without adding information and the contrasts represented by the standard deviations are unchanged. The standard deviations are trip-weighted like the averages.

The contrasts diminish to the left of the bold figures with averaging, most markedly between households and zones where there is averaging over a greater number of households per zone, compared with persons/household or trips/person.

Although there may be systematic differences in the variables between zones, some of the contrast will be due to random sampling. The contrasts arising from a very simple random process are shown in the bottom two rows of the table. They are based on dichotomous variables occurring randomly at the trip level, with either 50:50 or 25:75 probabilities; 75:25 would have the same effect as the latter. Standard deviations are calculated from the binomial theorem, assuming trips are divided equally between persons, trips and zones.

They correspond well with the rest of the dummy (not continuous) variables in the final column for the fully disaggregate case. Here the 'averages' are over one trip, so they are binary (0,1) variates. The

proportions of 0s and 1s determine both the mean and the standard deviation, so this contrast is 'absolutely uninformative' about any systematic variation.

With progressive aggregation to the zonal level, outward travel is underdispersed by comparison with the random sample. This can be explained by the natural pairing of outward and return trips. On the other hand, car availability and professional occupations are overdispersed, even after allowing for them occurring randomly at the household and person levels, rather than the trip level. This suggests systematic variations in them between zones, while males seem to be just about random.

Contrasts can reveal the influence of variables whether the contrasts arise from systematic or random effects. Good study design aims to minimise their reduction by averaging.

The strongest correlations between the variables are:

Personal income with household income	+.53
Personal income with professional occupation	+.40
Household income against only one worker	−.39
Service industry with professional occupation	+.31
Car sharing against ready car availability	−.30
Service industry against male gender	−.25
Household income with professional occupation	+.25
Age with no children	+.21

These are all plausible.

5.2 Modelling

Models were fitted to the HBW internal person trips by car. Apart from the exclusion of parking charges from costs, this is the dataset used by the WTSM and previous analyses of deterrence functions.

The base model had a simple Exponential deterrence function of cost. The variables were introduced as interactions with cost, cost.X where X is the variable. The main effect of X is absorbed into the production factors of the statistical model, or the trip generation stage of the transport model. This gives a deterrence function of the general form

$$\exp(-\lambda C_{ij} - \lambda_1 X_1 C_{ij} - \lambda_2 X_2 C_{ij} \ldots)$$

The variables were first added to the base model singly and then in combination all together. Finally a log(cost) term was introduced to the model, adding a $C_{ij}^{-\gamma}$ component into the deterrence function to represent the Tanner function and see if its fit is explained by any of the variables.

5.3 Results

5.3.1 Fitted coefficients

Table 5.4 shows the fitted coefficients λ_n.

Table 5.4 Fitted coefficients

Variable	Level of disaggregation			
	Zone	Household	Person	Trip
Cost	**0.06378**	**0.06378**	**0.06378**	**0.06378**
Individual				
Car available	−0.00440	**−0.00854**	−0.00854	−0.00854
One worker	0.01130	**0.00131**	0.00131	0.00131
No children	−0.01500	**0.00351**	0.00351	0.00351
Household income ×10⁻⁵	−0.00450	**0.00901**	0.00901	0.00901
Age ×10⁻²	−0.00500	0.01460	**0.00720**	0.00720
Male gender	−0.01590	−0.01954	**−0.01690**	−0.01690
Professional occupation	−0.02330	−0.01765	**−0.01368**	−0.01368
Service industry	−0.02680	−0.00507	**−0.00218**	−0.00218
Personal income ×10⁻⁵	0.00320	−0.04550	**−0.04040**	−0.04035
Outward	−0.02030	−0.01250	−0.01240	**−0.00276**
Peak	−0.05340	−0.03790	−0.03758	**−0.02624**
Share	0.01760	−0.00058	0.00148	**0.00220**
Combined				
Cost	**0.13754**	0.10872	0.10539	0.09487
Car available	−0.00388	**−0.00470**	−0.00468	−0.00492
One worker	−0.00564	**0.00586**	0.00303	0.00381
No children	−0.01261	**0.00332**	0.00344	0.00223
Household income ×10⁻⁵	−0.02410	**0.00710**	0.00330	0.00403
Age ×10⁻²	−0.01538	0.03051	**0.02669**	0.02861
Male gender	−0.04087	−0.01748	**−0.01461**	−0.01459
Professional occupation	−0.01623	−0.00899	**−0.00730**	−0.00751
Service industry	−0.02155	−0.00627	**−0.00453**	−0.00395
Personal income ×10⁻⁵	0.08910	−0.03950	**−0.03020**	−0.03192
Outward	0.00232	−0.00725	−0.00552	**−0.00064**
Peak	−0.04652	−0.03434	−0.03472	**−0.02393**
Share	0.00941	−0.00606	−0.00338	**−0.00146**

Bold – level of disaggregation at which the variable occurs

The top line shows the cost coefficient for a simple Exponential model. This does not vary with disaggregation across the table. It is the same as previously fitted to cost recalculated without parking charges in table 4.34.

The first of the two main blocks shows the coefficients of variables when entered individually. Coefficients are shown bold at the level of disaggregation where the variables occur. Replication of the variables with further disaggregation, to the right in the table, does not alter the coefficient.

This is not the case in the second main block of the table, where all variables are present in the model simultaneously. These models showed some signs of instability, as is to be expected when a lot of variables are entered into a complex model. At the top of this block, the main coefficient of cost acts as an intercept for continuous variables, extrapolated to zero incomes and age.

Other tabulated values are differences in the cost coefficient according to whether the variable is present or not. Coefficients for continuous variables have been scaled to represent the effects of having $100,000pa more income, or being 100 years older. Since the coefficient of the main effect of cost is positive, negative coefficients show a reduction in cost deterrence or a tendency to travel further. Several coefficients are a substantial proportion of the main cost coefficient.

Coefficients change in sign for many of the variables. Peak travel, service industry, professional occupation, male gender and full car availability are consistently associated with more travel, but there is a change of sign for car availability when a log(cost) term is introduced for the Tanner (not shown). Most changes appear between zonal and household aggregation.

5.3.2 Measures of fit

Table 5.5 shows the changes in deviance. The first two rows show the attraction balancing factor and the main cost effect, entered in that order. To the first decimal place, they are the same for all levels of disaggregation.

Table 5.5 Changes in deviances

Variable	Level of disaggregation			
	Zone	Household	Person	Trip
Attraction balancing	1313.43	1313.42	1313.44	1313.43
Cost	**2130.68**	2130.68	2130.68	2130.65
Car available	0.27	**4.68**	4.68	4.68
One worker	0.90	**0.09**	0.04	0.09
No children	2.86	**0.88**	0.85	0.88
Household income	0.27	**4.27**	4.27	4.27
Age	0.02	0.82	**0.25**	0.25
Male gender	1.33	15.38	**19.57**	19.62
Professional occupation	6.64	17.60	**13.11**	13.15
Service industry	5.75	1.05	**0.26**	0.26
Personal income	0.03	30.82	**30.49**	30.54
Outward	0.51	1.85	2.51	**0.55**
Peak	17.77	52.40	62.96	**43.94**
Share	2.15	0.01	0.10	**0.26**
Total	38.52	129.85	139.09	118.50
Combined	33.78	100.60	107.68	90.34
Log(cost) for Tanner	60.06	47.27	45.81	49.05

Variable	Level of disaggregation			
	Zone	**Household**	**Person**	**Trip**
Total residual deviance	4215.16	8104.42	8591.23	8688.87
Degrees of freedom	31,060	236,026	335,614	587,479
Mean residual	0.1357	0.0343	0.0256	0.0148
Expected residual	0.1393	0.0353	0.0274	0.0181

Bold – level of disaggregation at which the variable occurs

The next three sets of rows show the effect of variables entered individually. In each set, the level of disaggregation at which the variable occurs is shown in bold. Again, there is no change with further disaggregation to the right across the table. With more aggregation to the left, the change in deviance tends to diminish, particularly at the zonal level. However, there are exceptions to this trend, and not only in marginal values.

All changes in deviance are at least one order of magnitude, and usually two or more, smaller than the main effect of cost. Despite this, several variables appear significant. Taking a critical value around 4 (from χ^2_1), the only variables significant at all levels of aggregation are professional occupation and peak travel.

Peak travel is consistently the stronger of these, with larger coefficients. In this 24-hour distribution model, costs do not depend on time of day, so this is not simply an artefact of higher costs during the peak. With the current interest in peak spreading and advanced logit models of time choice, this may raise some chicken-and-egg issues. However, demand modelling by separate time periods, as in the Transport Model for Scotland, would accommodate this effect.

Professional occupation has a longer but chequered history in transport modelling as a white/blue collar subdivision of the work purpose.

Personal income and male gender appear strongly significant once disaggregated to household and beyond, but insignificant when aggregated to zone. The reduction in significance is quite disproportionate to the reduction in contrast due to averaging seen in table 5.3, or to the reduction in the number of cases. Because the deviances are for variables entered singly into the model, personal income is not being aliased with household income.

Household income and car availability are just significant when disaggregated to household, but again show little significance at the zonal level.

In contrast, service industry is significant at the zonal level but not at any greater disaggregation.

5.3.2.1 Total and combination

The changes in deviance from all the variables entered separately are totalled on the row of the table below them.

The row below that shows the change of deviance when all the variables are entered into the model simultaneously in combination. This is less than the total of the individuals because of the correlations between them, but still about 80% of the total, so there is not too much aliasing between the effects.

5.3.2.2 Tanner

The next row shows the change in deviance when entering a log(cost) term to form a Tanner deterrence function. The term is entered after all the variables combined, so a reduction from 66.1, the value with all the variables absent (and not showing in the table), indicates that the disaggregate variables can explain some of its fit.

Although this occurs to some extent, most of the effect of the log(cost) term remains and is larger than that of any of the variables, except for peak travel at household or person disaggregation. All 12 variables combined barely cause twice its change.

The Tanner function remains concave compared with the Exponential.

No interactions have been fitted between the variables and log(cost), or a Tannerised cost (section 4.2.4).

5.3.2.3 Residuals

Residual deviances from models with all variables in combination, but no log(cost) term, are shown at the bottom of table 5.5. The totals increase with disaggregation, but not as fast as the number of degrees of freedom, so the means decrease with sparsity. They are slightly smaller than the expected mean residual deviances and become more so with increasing sparsity from disaggregation.

5.4 Summary

Disaggregation can reveal effects on trip distribution of household, person and trip variables that are not apparent at the conventional zonal level of aggregation. The process is not regular and some effects appear significant only at the zonal level.

In Wellington, peak-period trips appear to be less deterred by cost at all levels of aggregation, as do those by professionals. This tendency to longer trips is only significantly associated with income and males in disaggregate models, and service industry only appears significant at zonal aggregation. All these significances as measured by change in deviance are small compared with the main effect of cost (separation), although the coefficients can be a substantial proportion of the cost coefficient. The Tanner function remains a significant improvement on the Exponential.

Change in deviance is invariant where disaggregation only involves replication of the same information with no further contrasts, and the numbers of replications are reflected in the weights of the aggregate. In these particular circumstances, conventional gravity model calibration at the zonal level is equivalent to disaggregate modelling at the household, person or trip level. All these are fitted by the same GLM formulation.

6 Fitting mixed logit models by hierarchical generalised linear model

6.1 Introduction

Choice occurs at all stages of the transport model, most obviously in the choice of mode (car, bus etc) but also in the choice of destination, route and time of day. The modelling of choice has been developed from economic concepts of random utility by McFadden and others, and applied to trip distribution models by Cochrane (1975, see section 2.1.3).

Analyses so far have allowed for a single level of randomness or uncertainty, common to all choices of destination. Choice modelling has developed to address different levels of uncertainty in different choices. In transport modelling, such differences are found between trip distribution and mode split, or between similar and dissimilar modes, as arise in the 'red bus, blue bus' paradox.

Mixed logit is at the forefront of choice modelling. Train (2003) has shown that it can fit any form of random utility model, including the nested logit, which is the most advanced form of choice model in regular use for transport models.

Current methods for fitting mixed logit models, such as Biogeme, require random simulation methods for integration. Even with modern processors, these are computationally demanding and can require considerable skill in setting up and interpreting the simulation process.

GLMs are extensions of least squares regression by Nelder and others. With a Poisson error and logarithmic link (log-linear), they can be formulated to fit multinomial logit models. This is the basic form of choice model – effectively an unmixed logit model. A wide variety of trip distribution models, of destination choice, have been fitted by GLMs in this study.

GLMs have been extended further to hierarchical GLMs (HGLMs) by Lee et al (2006). These can incorporate correlating error terms akin to some forms of mixed logit model. They use an iterative method built around GLMs which should be more efficient than simulation and has proven so in some applications.

At first sight, neither Train nor Nelder (pers comms) could see why HGLMs should not fit mixed logit models, possibly with a degree of approximation or some restriction of scope, but probably with increased computational efficiency. If HGLM could fit mixed logit models, it would bring the fitting of trip distribution models by GLM to the forefront of modelling practice. An HGLM has been used to fit Rasch models and the item response theory of psychological measurement by Kamata (2002) and Williams and Beretvas (2006), but it may deal in a hierarchy of fixed rather than random effects.

This chapter sets out the equivalence between the mixed logit model and HGLMs in mathematical terms. To demonstrate the hypothesis empirically, a dataset is generated on the framework of a mixed logit model for fitting by HGLM. Such datasets are also generated by random utility maximisation following the theory of choice, and nested logit following transport modelling practice.

The characteristics of the datasets were examined. In doing so, it became apparent late in the study that the standard dataset, already used for much analysis, lacked power of distinction. Larger datasets posed computational problems.

Mixed logit has been approached by its alternative interpretations of error components and random coefficients. Fewer, larger groups with tastes in common, differing only between groups, were tried instead of the usual disaggregate approach, with tastes varying between every decision.

Generalised linear mixed models (GLMMs), a subset of HGLMs with normal random effects, have been run to fit random coefficient models and for comparison. Cross-checks have been made by simulation methods (using the Biogeme software package), and other approaches are discussed.

6.2 Hypothesis

The hypothesis is that mixed logit models and HGLMs can be equivalent. This is shown in the terms of Train for the mixed logit model. HGLMs are described in the terms of Lee et al and their implementation in Genstat.

6.2.1 Mixed logit

Train (2003, section 6.1 p139) defines a mixed logit model as

$$P_{ni} = \int L_{ni}(\beta) f(\beta)\, d\beta$$

where $L_{ni}(\beta)$ is the logit function

$$L_{ni}(\beta) = \exp(V_{ni}(\beta)) \,/\, \Sigma_j \exp(V_{nj}(\beta)),$$

and $f(\beta)$ is the mixing distribution; in plain logit, β takes a fixed value.

In choice modelling, P_{ni} is the probability of individual n choosing option i from among j=1..J options available.

6.2.2 Individual decision

'Individuals' are sets of alternative options from which one is chosen. In observed data (revealed preference), each choice is typically made by a different person, who can be thought of as the individual. In panel surveys (stated preference) each person may be given several choices to make; the 'individual' is then the case from which a single choice is made; several individual cases may be put before a single person. Coefficients may then be specified to vary only between persons, or within-person, eg learning or response fatigue.

The individual decision has its own value of β_n ('knows the values of his own β_n' in Train (2003, section 6.2, p141) in the context of random coefficients). Therefore, the individual decision can be treated as a simple logit model with fixed β_n.

6.2.2.1 Multinomial and Poisson distributions

Sampling with the probabilities of a logit model gives a multinomial distribution. A multinomial distribution is equivalent to a Poisson distribution for each of the possible outcomes under the constraint of the total for all outcomes (Sen and Smith 1995, section 5.2.4; McCullagh and Nelder 1989, sections 5.2 & 6.4; Lee et al 2006, p48).

6.2.2.2 Generalised linear model

If $V_{ni}(\beta)$ is linear in parameters β

$$V_{ni}(\beta) = \beta x_{ni} \qquad \text{Train (2003, equation 6.1)}$$

(or $-\lambda$cost, for trip distribution with Exponential deterrence)

A GLM can then fit the multinomial logit model. Each option presented to an individual is a unit of data. The GLM is specified with a Poisson distribution and a logarithmic link to a linear predictor $\eta = \beta x_{ni}$. The denominator $\Sigma_j \beta x_{nj}$ of the logit function is fitted by a constant common to all options j open to individual n. This provides the constraint on the total for all outcomes.

This established method of fitting a multinomial logit model is applicable to an individual because β_n is fixed for all options presented to the individual n.

6.2.3 Population of decisions

When fitting a mixed logit model to a population, β_n generally varies between individuals and cannot be observed directly by the analyst.

The linear predictor or utility βx_{ni} can be decomposed into fixed terms, whose parameters do not vary across the population, and random terms representing the variation between individuals. In his discussion of error components Train (2003, p143) writes this as

$$\alpha x_{nj} + \mu_n z_{nj} + \varepsilon_{nj}$$

fixed random GEV/logit

where the third term ε is a generalised extreme value (GEV) error that leads to the logit probabilities under utility maximisation.

In a transport choice model, x might be a travel cost which deters all individuals to a common degree, α. Under a random coefficient model, z can be the same cost, but its effect on an individual varies from the common value α by μ_n. Under an error component model, z can be a set of dummy variables that define a nesting structure; μ_n can represent a commonality between the red bus and the blue bus.

6.2.3.1 Hierarchical generalised linear model

Hierarchical GLMs (HGLMs) can incorporate a random term in the linear predictor of a GLM. Quoting Lee and Nelder (1996), Lee et al (equation 6.1) write the linear predictor as

$$g(\mu) = \eta = X\beta + Zv$$

where μ is the mean response and v=v(u) is a monotone function of random component u.

This is equivalent to the mixed model as formulated above:

Table 6.1 Comparison of mixed model and HGLM components

Component	Mixed model	HGLM
Fixed terms; design matrix	x_{nj}	X
Fixed coefficients	α	β
Error components; random terms	z_{nj}	Z
Random coefficients	μ_n	v
GEV term/logit form	ε_{nj}	Logarithmic link g()

Generalised linear mixed models (GLMMs) also take this form when v is normally distributed.

The statistical software package Genstat includes algorithms to fit such models.

The constants to fit the denominator of the logit function, and constrain the model to the total of outcomes for an individual, become a factor with one level for every individual in Genstat terminology.

6.3 Data set generation

To test the hypothesis empirically, datasets have been generated with known parameters. Fitting an HGLM to the data should then recover the parameters as fitted coefficients. Datasets were prepared in Genstat,

using its intrinsic pseudo-random number generation procedures. Notation generally follows Genstat, so for an HGLM

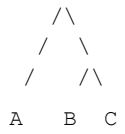
$$\text{probability } \mu = \exp(X\beta + Zv + \text{balance})$$

or in terms of choice theory

$$\text{utility} = X\beta + Zv + \varepsilon_{nj}$$

6.3.1 Structure

All data sets had three options, with a common error component correlating two of them. This is equivalent to the simplest nested logit model, with the two correlated options in a separate nest below. Options B and C can be seen as the red and blue buses of the classic transport dilemma.



There is only one attribute that systematically affects choice (utility X below), so, with a single extra random component (or nesting ratio) to fit this is a simplest case of mixed logit.

6.3.2 Design

6.3.2.1 Fixed attributes, X

To avoid negative coefficients, the systematic component (X) is defined as utility rather than its opposite, cost. For the standard dataset, X is taken as all combinations of -1, 0 and 1 for the options A, B and C, a full factorial design, generating 27 combinations:

Case	A	B	C
1	-1,	-1,	-1;
2	-1,	-1,	0;
3	-1,	-1,	1;
4	-1,	0,	-1; and so on
:	:	:	: up to
27	1,	1,	1.

This pattern is repeated 10 times, giving 270 cases in the standard dataset. In the absence of correlations, the probabilities of choosing A, B or C will be equal over the whole dataset.

6.3.2.2 Random error structure, Z

The structure of the correlations between options is described by the factor Z. This takes two levels for each case; one for option A and the other common to options B and C. The error component v takes a different random value for each level of Z; taking a common value between options B and C for a case produces a correlation between them.

Case	Option	Levels of Z
1	A	1 0 0 0 0 …
1	B	0 1 0 0 0 …
1	C	0 1 0 0 0 …
2	A	0 0 1 0 0 …

```
2       B       0 0 0 1 0 …
2       C       0 0 0 1 0 …
3       A       0 0 0 0 1 …
:       :       :
```

6.3.2.3 Balance

This third term represents the denominator of the logit function and ensures that the probabilities of all the options for a case sum to 1. It is generated from the reciprocal of the sum of the unbalanced probabilities, taking a common value for all the options in a case:

$$\exp(\text{balance}) = 1 \;/\; \Sigma_{\text{options}} \exp(X\beta + Zv)$$

In a GLM or HGLM it is fitted as the coefficient of case, a factor with one level per case.

6.3.3 Coefficients

6.3.3.1 Fixed coefficient, β

The probabilities for particular combinations of utilities X are determined by their coefficient β (adopting Lee's notation for HGLMs; this is α in Train's notation). This parameter has been taken as 1 in the standard datasets. The resulting probabilities are shown in figure 6.1.

With β taken as 1, the set of Xs gives a good spread of probabilities. It is the contrasts between the probabilities that provides the information to recover β in model fitting, so the design looks as though it has a reasonable power of detection for β. Work done by Toner et al (1999) on the design of stated preference questions suggests that more points at the edges and less at the centre may improve the power further.

6.3.3.2 Random coefficient, σ, for realisations of v

For each level of Z, the correlating/mixing error component v takes a different value drawn at random from its distribution. This distribution is taken to be normal to allow alternative methods to be used, in particular GLMMs. The gamma distribution is conjugate – see section 6.11.9.

The scale of v is described by its standard deviation σ, which is the key parameter in fitting the mixed model. Its standard value was taken as 0.5, in order to be smaller but still substantial compared with the overall error ε implicit in the logit model. This has a Gumbel distribution with standard deviation $\pi/\sqrt{6} = 1.28$.

In later runs, σ was taken as 2 for greater power of detection, still in keeping with the values from practical models below. This value had also been used in some preliminary trials.

6.3.3.3 Observed values of σ

This section considers realistic values of σ for generating test data. No major transportation model is known to be built on mixed logit, so there are no values of σ calibrated directly from observations. Ratios of fixed coefficients (βs) for different levels of nesting are considered here as a proxy based on the calculations for nesting models in section 6.3.5.3.

Bly et al (2001) assembled choice coefficients from a number of transport models. Table 8.8 in Bly et al gives ratios between coefficients for mode split and those for trip distribution. For home-based work, there are seven diverse values ranging from 0.574 to 3.00 with a mean of 1.28. The coefficients forming these ratios may have been estimated in separate models, or in different segments of a joint model, as in the WTSM and not in a formal nested model.

A nested model was formally fitted in the West Midlands model PRISM by RAND (2004, task 1, table 25). The ratio of coefficients is referred to as the structural parameter (θ). For the main commuter model it is estimated as 2.54 or 0.48 (=1/2.08) for alternate orderings of mode split and distribution. A t value of 21

suggests that a much smaller value could be detected, but the sample size and structure are not readily determined. Similar structural parameters were fitted for other purposes.

Analysis of HBW trips in the WTSM did not reveal a clear ordering of mode split and distribution and a simultaneous model was fitted. There was complex segmentation by mode, household car availability and spatial divisions, resulting in a large number of coefficients. The greatest range is between:

0.018 by car within urban sectors, and

0.1175 by public transport to the CBD

both for households with some cars, but more adults. This is a ratio of 6.5.

More compatible comparisons of car and public modes can be calculated from coefficients for 'other' geographic segments, generally longer distance but not to the CBD, as

0.0445/0.0215 = 2.1 for households with more adults than cars, or

0.0391/0.0273 = 1.4 for households with at least one car per adult.

The 1991 London model had a similar structure. Table 13.2 'Final deterrence parameters for white collar work' in MVA (1998) gives coefficients as:

0.0727 Car, central London destination

0.0325 Public transport, some car owning, central London destination, ratio = 2.2

0.0378 Public transport, non-car owning, central London destination, ratio = 1.9.

Ratios for other purposes and destinations are similar.

The table in section 8.3.3 of Bly et al (2001) compares distribution deterrence coefficients between car and public transport modes, from three transport studies. The ratios are generally greater than 2, including all those for home-based work.

These all show that ratios of 2 or more between choice coefficients are frequently found in transport models. Entering this value for θ in table 6.2 or the equations in section 6.3.5.3 gives a value of greater than 2 for σ.

6.3.3.4 Randomness of utility, ε

The underlying random term in utility ε was drawn from a Gumbel distribution for each case and option. Only the random utility maximisation (RUM) method (see section 6.3.5.1) uses it in calculations; other methods use the logit form which it leads to under RUM theory.

6.3.4 Output/response variables

In generating a dataset, three dependent variables appear. In the long-run of repeated randomisation their averages are the same and sum to 1 over the options for each case.

6.3.4.1 Probability, μ

This includes the systematic component $X\beta$, the random component Zv and the normalising factor, balance. With balance included, the probabilities of all the options for a case sum to 1. Probability is a continuous variable.

This probability is taken as the mean for overall randomisation, either to give Y values for each of the options or to sample one choice from the options.

6.3.4.2 Unconstrained outcome Y

Each option in a case is randomised independently, so the total is not constrained to 1. This is the usual form for an HGLM. Several alternative distributions can fitted by an HGLM, as specified in the `DISTRIBUTION` option of `HGFIXEDMODEL`. If the distribution is normal, the data can also be fitted by REML.

Y can be continuous or discrete, according to the characteristics of the overall randomising distribution. For the nearest equivalent to a mixed logit model, the distribution is Poisson.

6.3.4.3 Choice

In a mixed logit model, one and only one option is chosen. A variable, choice, is set at 1 for the chosen option and zero for others (so the sum is always 1). It has a multinomial distribution with the probabilities μ given above. Choice is a discrete variable.

Unmixed logit models of choice without the Zv random component can be fitted by plain GLM.

6.3.5 Methods

Three different methods can be used to generate datasets: RUM, mixed logit and nested logit.

6.3.5.1 Random utility maximisation

This follows the mechanism of choice theory. Random values are drawn for both v and ε, and both are added to the systematic term $X\beta$ to give the random utility. The option with the highest of these utilities is then chosen.

The result is an output for choice. To find the underlying probability, a large number of randomisations have to be averaged.

6.3.5.2 Mixed logit

In this method, only the error component v is randomised. It is added to the systematic term $X\beta$ and the resulting utility is used to calculate the probability for each option using the logit function. In effect, the realisations of the randomised v are incorporated into the systematic effect; the effects of maximising with uncertainty ε are represented by the logit function, as derived from the RUM theory.

The resulting probabilities are for given draws of v, so again they have to be averaged over many randomisations of v to find the underlying probabilities for given combinations of X. However, there is less randomness than under the RUM method, because there is no binary (0,1) sampling of choice to average out.

On the other hand, choice still has to be generated from the probabilities. A random draw was taken from a uniform 0-1 distribution; the corresponding option on the cumulative probability of options was chosen. (Law and Kelton 1991). This ensures that only one option is chosen.

Other outcomes (Ys) can be generated from the probabilities of individual options. They can be taken as the means of Poisson or normal distributions from which draws are taken. Since the draws are independent they will not necessarily sum to 1 over the options for a case, although their expectation will.

This corresponds with the formulation of an HGLM, with the logit function provided by a logarithmic link. Mixed logit was the main method used to generate data for HGLM fitting.

6.3.5.3 Nested logit

This is an approximation to the other two methods, but yields analytical results without recourse to randomisation.

In the upper nest (A v BC), the error component v in which A differs from B or C is added to the overall randomness of ε. With a GEV Gumbel distribution, ε has a variance of $\pi^2/6$, so adding the variance of v, σ^2, gives an increased variance of

$$\pi^2/6 + \sigma^2$$

or a relative increase on the linear scale of

$$\theta = \sqrt{(1 + \sigma^2/(\pi^2/6))}$$

With this greater randomness in the upper nest, systematic effects are less pronounced, so the effective systematic coefficient is reduced to β/θ. This nesting ratio θ is tabulated against σ in tables 6.2 and 6.3.

In the lower nest (B v C), the error component v is common to B and C, so ε is the only random component of error between them, and the systematic coefficient is β. The relative probabilities between B and C are determined by differences in their systematic X attributes; they are not affected by common shifts of origin, as is the effect of v.

With these systematic coefficients for the upper and lower nests, the probability of each option can be calculated analytically. Random draws again have to be taken from the cumulative probabilities of options to give a choice, or other outcome Y values can be drawn independently from distributions with the probabilities as means. In this method the error component v is incorporated into overall random term (ε), whereas in the previous mixed method it was incorporated into the systematic term (βX).

Two approximations are made in this method:

- The combined or inclusive utility of the lower BC nest is calculated by the logsum with the coefficient β, in accordance with the RUM theory. This theory is based on the randomness ε in the individual utilities of B and C, but it also predicts that the combined utility has randomness on the same scale. However, for calculations in the upper nest, the same basic theory requires the randomness to be greater by θ.
- The same basic theory requires the randomness to take a Gumbel GEV distribution. Mixing a normal distribution for v with a Gumbel distribution for ε does not give another Gumbel distribution for the upper nest. As v becomes large compared with ε, the upper nest choice changes from logit to probit; there is little practical difference between the two.

6.3.5.4 Control of randomisation

When generating datasets to compare methods, the same random draws are used for each dataset as far as possible.

The same random draws of v are used in both RUM and mixed logit methods. There are two draws per case; one for option A and the other common to options B and C.

The same random draws of a uniform 0-1 distribution are used for generating choice from the probabilities of options in the mixed and nested methods. There is one draw per case.

In the RUM method, there is a random draw of ε, one for each of the three options in a case. These also lead to the generation of choice, but no way could be seen to give consistency with the uniform distribution drawn for the mixed and nested methods.

For independent randomisations of Y, there is a uniform 0-1 draw for each of the three options in a case. This is then converted to the desired distribution by its cumulative function. This gives consistency in that if, say, a particular option is randomised high under a Poisson distribution, it will appear similarly high under a normal distribution.

Seeding of these draws was controlled to ensure reproducibility, and where desirable consistency, between runs.

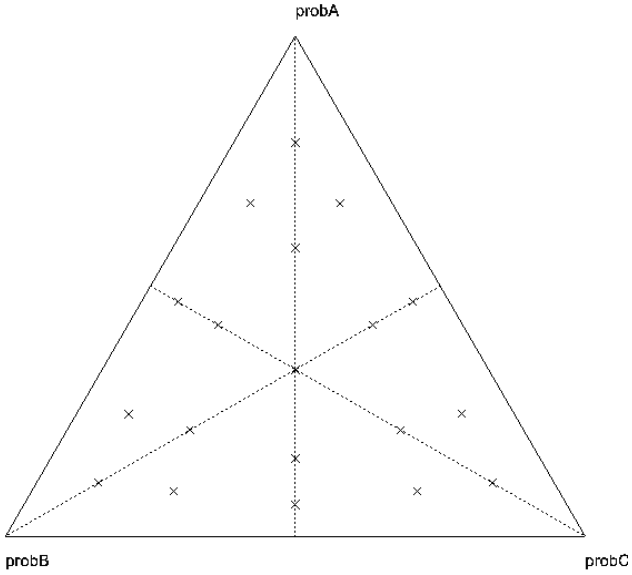
These 'best-practices' evolved in the course of the study and were not all used in early runs when problems were encountered with some Genstat randomisation functions.

6.4 Characteristics of dataset

6.4.1 Systematic effects

The probabilities resulting from the utilities X and their coefficients β are shown in figure 6.1.

Figure 6.1 Systematic probabilities (β=1)



The uppermost point in figure 6.1 represents a highest probability of choosing A and a low probability of choosing B or C. This occurs when the utility of A is high x_A=1, and the utilities of B and C are low x_B=x_C=-1. Using the logit formula:

Probability of A = $\exp(\beta x_A) / (\exp(\beta x_A) + \exp(\beta x_B) + \exp(\beta x_C))$

= exp(1.1) / (exp(1.1) + exp(1.-1) + exp(1.-1))

= e / (e+2/e)

= 0.787

Probability of B, or C = (1-Probability of A) / 2

= 0.106

The central point represents three combinations of Xs which are all equal:

```
 A  B  C
-1,-1,-1;
 0, 0, 0; and
+1,+1,+1.
```

Each of the six points in the inner ring represents two combinations of Xs, with the same differentials. For example, the point above the centre represents the combinations:

```
 A  B  C
+1  0, 0; and
 0,-1,-1.
```

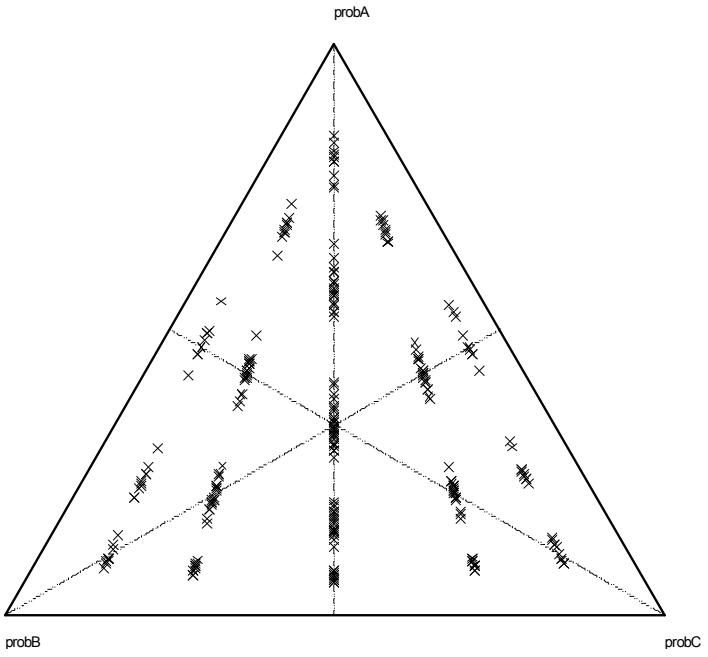
The 12 points on the outer ring represent one combination of costs each. Thus all the points represent the 27 combinations of Xs:

$$1\times3 + 6\times2 + 12\times1 = 27$$

As β is increased, the points in figure 6.1 will spread out towards the edges; as β is reduced, they will contract towards the centre (eg figure 6.4). The same effect will be produced by scaling all Xs up or down, but adding a constant to all Xs will not affect the relative probabilities.

6.4.2 Mixed logit randomisation

Figure 6.2 Probabilities after randomisation of mixing term



Individual probabilities of 270 cases (not aggregated); σ = 0.1

Figure 6.2 shows the effect of adding the random error component to the systematic effects of figure 6.1. Two values of v are drawn for each case, one for option A and one for B and C. After these are added to the systematic components, the probabilities are calculated by the logit formula.

For the purposes of the figure, the scaling of v, σ, is reduced to 0.1, limiting the spread of points around the systematic probabilities seen in figure 6.1. With the standard value of σ = 0.5 the groups overlap.

The groups can be seen to spread along tracks converging on the top apex. This is because the random error component is the same for options B and C, and therefore does not affect their relative probabilities.

6.4.3 Average probabilities

Figure 6.3 Average probabilities by different methods

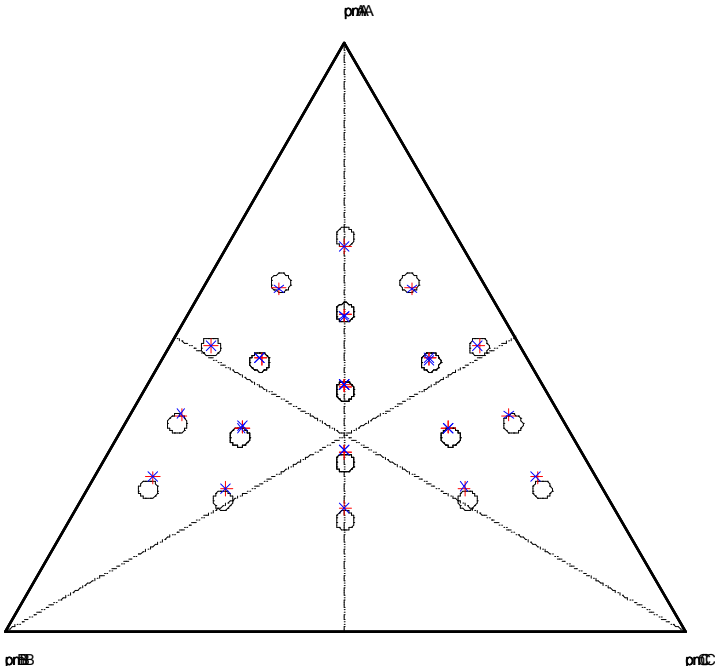


Figure 6.3 shows the long-run probabilities calculated by the three different methods:

Black O: Nested, calculated using logsum costs. Purely analytical, thus no sampling error.

Red +: Mixed. Random term randomised, then probabilities calculated by logit. Thus error from sampling the mixing distribution only.

Blue x: GEV. Random term and GEV term randomised; maximum utility selected. Thus error from both mixing and Poisson sampling.

A large dataset of 2,700,000 cases was generated, leaving average errors 1/100th of those in the standard set of 270. The cases have been aggregated by the combinations of X values for the options.

For the purposes of this figure, σ is increased to 2, to exaggerate the shifts from the systematic probabilities seen in figure 6.1. These shifts are along same converging tracks as seen in figure 6.2.

Compared with figure 6.1, the higher probabilities of A at the top of the plot are reduced. This seems contrary to an expected increase due to correlation in utility between B and C, representing a lack of choice between them.

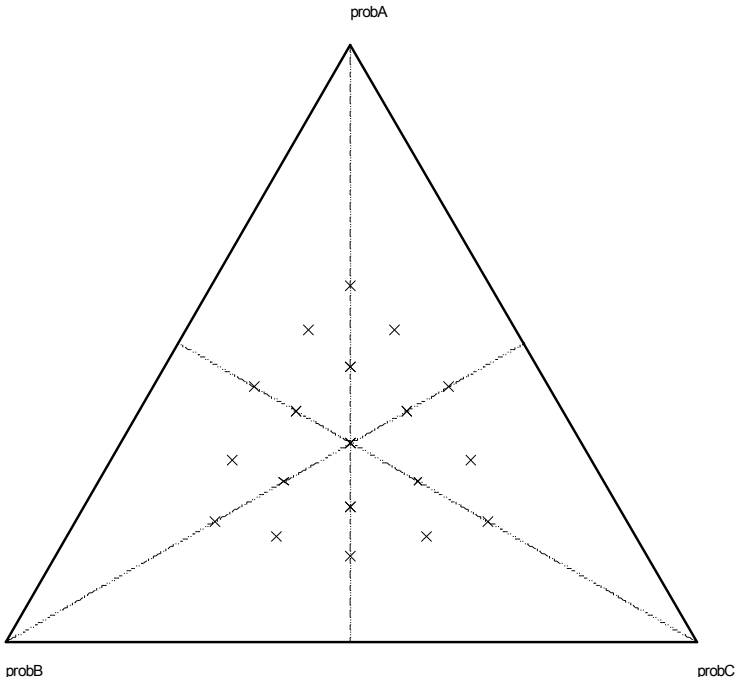
Upward randomisation of A's relative utility can only increase the probability of A by a limited amount; it cannot go over 100%. There is much more scope for a decrease in the probability with a downward randomisation, so the overall effect of randomisation is a decrease in probability. This is a consequence of non-linearity at the top of the logit curve and can be seen as a form of 'aggregation bias' (Ortuzar and Willumsen 1994, section 9.2; Train 2003, fig 2.1) when aggregating over the random error component.

There is an increased probability of A over the whole dataset because there is also an upward bias of the lower probabilities at the bottom of the plots.

This overall shrinkage on the vertical scale can be seen as a consequence of the random (mixing) component v increasing the underlying randomness ε in the utility maximisation model. The relative importance of the systematic utility is reduced so its coefficient β becomes smaller as for the upper nest in a nested model.

Figure 6.1 shows the effects of systematic utility with a coefficient $\beta = 1$, the value for the lower nest. Figure 6.4 shows them with a coefficient β set to$1/\theta$, the corresponding value for the upper nest.

Figure 6.4 Systematic probabilities scaled by upper nest coefficient



$\beta = 1/1.8525$ (corresponding to $\sigma = 2$)

Without any allowance for correlation between B and C, shrinkage is the same horizontally and vertically compared with figure 6.1. All points in figure 6.3 show an increased probability of choosing A compared with this plot.

6.4.4 Proportions choosing option A

Table 6.2 shows the average proportions of choice A for different levels of σ, the standard deviation of the correlating random component common to choices B and C. The table also shows the corresponding nesting ratio θ between the systematic coefficients (βs) for the upper and lower nest, calculated as in section 6.3.5.3.

The proportions are generated by three different methods: nested logit, mixed logit and RUM. Generation by mixed logit and random utility maximisation involve randomisation, which was kept to a minimum. The same set of 2,700,000 cases was generated as for figure 6.3, again leaving average errors 1/100th of those in the standard set of 270.

Table 6.2 Proportions choosing option A

Parameter correlating B & C		Generation method		
Mixing, σ	Nesting, θ	Nesting	Mixing	Random utility max
0	1	0.333333 by definition		
0.1	1.003	0.33365	0.33377	0.33370
0.5	1.073	0.34078	0.34327	0.34385
1	1.268	0.35825	0.36472	0.36520
2	1.852	0.39513	0.40423	0.40473
3	2.544	0.42080	0.42899	0.42964

Figure 6.5 Proportions choosing option A

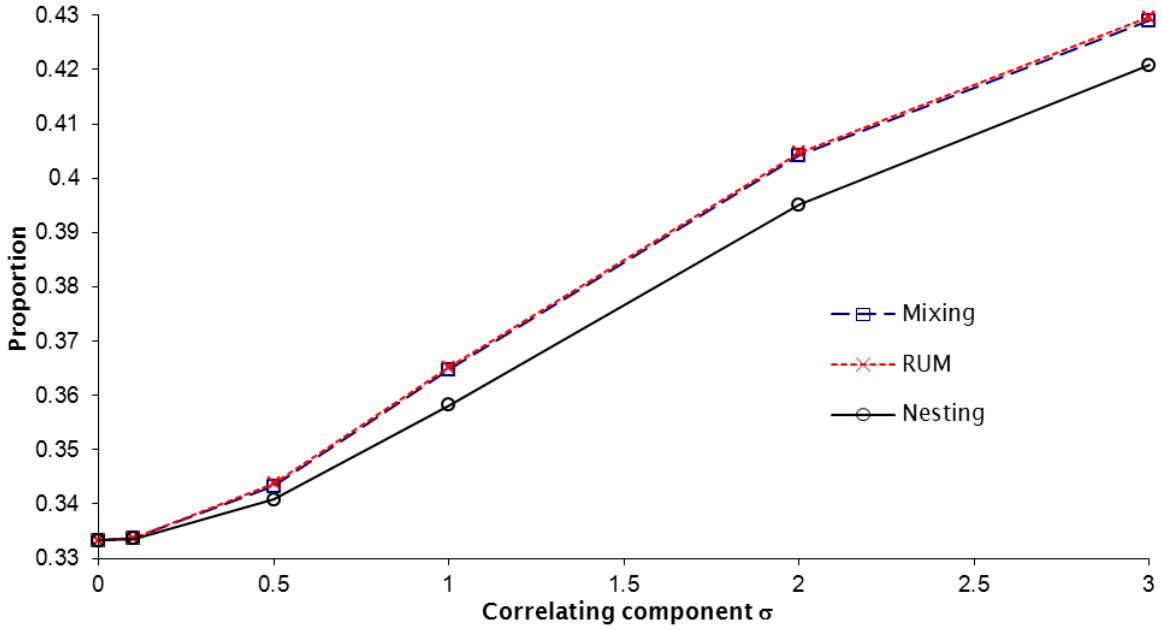


Figure 6.5 plots these values. When σ is zero, it has no effect on the proportion of A, which is 1/3, equal to B and C. As σ increases, mixed logit and RUM give very similar results despite randomisation. However, the analytic nested logit is consistently lower. This may be due in part to the 'impure' mixtures of normal and Gumbel distributions in the randomisations, where a pure Gumbel is presumed in simple logit choice theory. However, the nesting equations assume that the cost coefficient for the upper nest is β/θ, reflecting uncertainty in the joint/inclusive utility of BC combined; but the basic logit choice model shows that when B and C are combined in the lower nest with cost coefficient β, the uncertainty in the outcome is still represented by β, not β/θ.

6.4.4.1 Sampling error

A simple sampling error in the proportion choosing A can be estimated from a binomial distribution with $p = 1/3$, $q = 2/3$, $n = 270$. The standard error is then $\sqrt{(pq/n)} = 0.0286$. This is probably conservative, since the design of data X gives a range of probabilities whose mean is 1/3, even excluding mixing effects.

6.4.4.2 Power of detection

With this underlying sampling error of 0.0286 in the proportion of A, it will be difficult to fit a mixing component with $\sigma = 0.5$, whose effect on the proportion of A is only 0.01. The sample is unlikely to have the power to detect correlating/mixing components whose effect on the proportion of A is no greater than that of sampling error.

This approach to the power of detection depends on the proportion choosing A being a sufficient statistic, summarising all the information available about σ. However, there may also be information in the 'aspect ratio' of figure 6.3 – the greater shrinkage on the vertical scale than on the horizontal. This is a relatively large effect compared with overall upward shift, and could be seen as a contrast between the coefficients of utility (βs) for the upper and lower nests in a nested model. If this contributes further information about σ, the data can have greater power.

6.4.4.3 Sample size

Table 6.3 gives adequate sample sizes based on the sampling error estimation in section 6.4.4.1 and the effects given in table 6.2.

Table 6.3 Sample size

Parameter correlating B & C		Sample size	
Mixing, σ	Nesting, θ	cases	× standard sample
0	1.000	∞	∞
0.1	1.003	18,646,932	69,062.7
0.5	1.073	36,010	133.4
1	1.268	3609	13.4
2	1.852	707	2.6
3	2.544	389	1.4

These sample sizes are designed to give an effect four times the standard error. The effects are taken from the mixing method of generation; the RUM method is very similar, but the smaller effects from the nesting method require an even larger sample.

The sample size is given both as the number of cases and as a multiplier of the standard sample size of 270 cases.

A very large increase in the size of the dataset is needed for the standard value of $\sigma = 0.5$. A value of $\sigma = 2$ was chosen as still being within the empirical range of values for θ considered in section 6.3.3.3, and not becoming too large compared with the overall error implicit in the logit model ε with its standard deviation of $\pi/\sqrt{6} = 1.28$.

For $\sigma = 2$, table 6.3 suggests increasing the standard sample by 2.6. Attempts to fit HGLMs to datasets four or three times the standard size of 270 cases failed due to lack of space. (1GB RAM; there were indications that Genstat and/or Windows XP were failing to access the paging file on the hard drive.) With double the standard size of dataset, 540 cases, HGLMs took 10 hours for 99 cycles when they failed to converge. The designs of the larger datasets were simply replications of the attribute Xs described in section 6.3.2.1.

This approach to power of detection and adequate sample size was only found late in the study after much work had been done on standard datasets with limited powers of detection. This is described below; a few key re-analyses with the doubled sample size and $\sigma = 2$ have not revealed any significant

differences. Some preliminary runs that were made with $\sigma = 2$ but with the standard size had not met with any greater success.

6.5 Fitting mixed logit with HGLM

6.5.1 Base

The base HGLM to fit the dataset was coded in Genstat by:

```
HGRANDOM [DISTRIBUTION=normal; LINK=identity] Z
```

to specify v, the normally distributed random coefficient of Z

```
HGFIXED [DISTRIBUTION=poisson; LINK=log; DISPERSION=1] Case,X
```

to specify a balancing factor for each case, and β, the fixed coefficient of X, with a logarithmic link for the logit form and Poisson sampling, and

```
HGANALYSE Choice
```

to fit to choice.

Other settings were generally taken at defaults. Higher order Laplace transforms failed, so were not specified for these runs (modified code allowed them to be specified in later runs). Variations in the Aitken adjustments to assist convergence (`EMETHOD`) were tried. They appeared to affect the pattern of iteration, but to no clear advantage.

Because the mixing distribution is normal, the same model can be fitted as a GLMM, specified by:

```
GLMM [DISTRIBUTION=poisson; LINK=log; DISPERSION=1; RANDOM=Z;
                              FIXED=X+Case; ABSORB=Case] Choice
```

Only one randomisation of the data was generated. Multiple randomisations are time consuming because Genstat stops when HGLM fails to converge. Problems were encountered controlling the seeding of Poisson randomisation. More importantly, it was felt that a larger scale of computing and analysis would be undesirable because process would oust comprehension. If an analysis is correct and robust, it will appear successful in any one case; there should be no need to try many.

There was no variation in the number of cases, 270. This number was thought sufficient to reach a solution. It leaves substantial sampling error, but a much larger number of cases was needed to reduce this, which would be time consuming where convergence was slow or not completed. This sample size gave reasonable run times and turnarounds.

There was no variation in the number of options, three. This is a minimal case, and hence the most comprehensible; there is no obvious benefit from greater complexity.

6.5.2 Variations

From this base case, a number of variations were tried.

Regression on Y instead of choice. This is closer to the usual form of a GLM as there is no constraint on Y totalling 1 across the options for a case, as there is for choice.

Selecting cases where s**um(Y)>0** or s**um(Y)=1**. These should still be valid datasets, since they are drawn at random about a mean of 1 for the total of every case; the process is absorbed in case, and cannot be biased in X or Z.

Aggregation of cases with the same set of X values (utilities) for options. Each resulting case is thus the outcome of (say) 10 choices, rather than one, multinomially distributed, avoiding binary outcomes. This implies some averaging of the random effects v. If the number of initial cases that are aggregated were increased, randomisation would be averaged out and the outcomes would tend to the underlying probabilities, as in figure 6.3 and table 6.2.

DISP=* instead of 1. A fixed dispersion is justified by knowing there is a pure Poisson process. Using that information may help the fit of other coefficients.

β = 2 or 0.5 instead of 1. From the spread of resulting probabilities in table 6.2, $\beta = 1$ and the design matrix of X seem to provide reasonable contrasts.

σ = 2 or 1 instead of 0.5. 0.5 gives a variance substantially less than that of the Gumbel underlying random utility and leading to the logit form without being trivial. It may still be rather small to show significance from a modest number of cases.

Randomised X, instead of combinations of -1,0,1 for options A,B,C. This was to check that the regular data structure was not causing problems. X was randomised N(0,1)

6.5.3 Results for the standard dataset

Fitting a simple log-linear GLM with Z as a fixed effect appears to recover the fixed coefficient β. Half the coefficients of Z are aliased, and many remaining values fitted to both it and case take extreme values.

The results for fitting GLMM and HGLM are shown in table 6.4.

Table 6.4 Output from variations – standard dataset

Variation	GLMM					HGLM				
	Cycles	β		σ^2		Cycles	β		σ^2	
		X	se	Z	se		X	se	Z	t
Expectation		**1**		**0.25**			**1**		**0.25**	
Base	2	1.038	0.1172	0	bound	99				
Regress Y	4	0.9922	0.11406	0.091	0.157	16*	0.992	0.1	0.09402	-4.89
Sum Y > 0	4	0.9922	0.11406	0.091	0.157	16	0.9922	0.114	0.0913	-4.89
Sum Y = 1	2	1.085	0.1935	0	bound	20	1.0847	0.193	0.000465	-0.69
Aggregate	2	1.038	0.1172	0	bound	20	1.038	0.1171	0.000223	-0.87
Disp =*	20	1.041	0.1099	0.5354	0.2604	16	1.037	0.112	0.3283	-4.47
σ = 1	4	0.92	0.11262	*0.103*	*0.226*	57	0.92	0.113	*0.1049*	*-5.1*
σ = 2	5	0.7521	0.10784	*0.305*	*0.235*	99				
β = 0.5	2	*0.4756*	*0.09655*	0	bound	99				
β = 2	4	*1.844*	*0.1747*	0.11	0.277	20	*1.844*	*0.174*	0.115	-4.38
Random X	4	0.975	0.10651	0.098	0.243	75	0.9751	0.106	0.09091	-4.87

Italics: not expected to match the expectation for the base model in the top line

* matrix h not positive semi-definite while executing the CHOLESKI function; no likelihood statistics

Fixed coefficient, β

This appears to be reproduced, including variations in its value, $\beta = 0.5$ or 2. However, an alternative value, for the upper nest in a nested logit model, differs by only 7% for $\sigma = 0.5$.

Random coefficient, σ

This does not appear to be reproduced, except possibly for Disp =* with HGLM. Otherwise values are low, or on bounds.

GLMM v HGLM

There is remarkable consistency for two independent algorithms. (Roger Payne of VSN (pers comm) seemed to suggest that GLMM now used the HGLM algorithm, but procedure source codes in Genstat 9.1.0.147 appear to differ).

Cycles

GLMM converges faster and always offers solutions, and usually has σ^2 on a bound if HGLM fails to converge or estimates very low values. There is an exception to GLMM's faster convergence when Disp =*.

Accuracies

Standard errors are generally too high to distinguish recovery of the coefficients. The accuracy for σ^2 in HGLM is given as the output t statistic for lambda Z, the log of σ^2. Crudely applied, this suggests that HGLM offers higher accuracies than GLMM for the random term.

6.5.4 Results for the doubled dataset

The main variants were re-run on a larger, more powerful dataset with σ set to 2 rather than 0.5. Higher order Laplace transforms were used `[MLAPLACE=1; DLAPLACE=2]`, as modified code for them had become available. The results are shown in table 6.5.

Table 6.5 Output from variations – doubled dataset, σ = 2

Variation	GLMM					HGLM				
	Cycles	β		σ^2		Cycles	β		σ^2	
		X	se	Z	se		X	se	Z	t
Expectation		**1**		**4**			**1**		**4**	
Base	7	0.7002	0.07523	0.348	0.167	99	failed to converge			
Regress Y	9	out of bounds		0.788	*	21	failed – zero response variate			
Sum Y > 0	10	0.7769	0.08164	0.788	0.180	99	failed to converge			
Sum Y = 1	1	0.6198	0.11848	0.277	0.264	12	failed – zero variance function			
Aggregate	4	0.7196	0.08117	0.072	0.054	12	0.7189	0.0809	0.06539	-5.96
Disp =*	17	0.8103	0.07151	1.7014	0.2372	50	0.8196	0.089	1.463	4.32
Random X	6	0.7130	0.06782	0.442	0.177					

This does not show any improvement in fit to the more powerful dataset. The fixed coefficients β are now distinctly below their expected value of 1, but still above the value for the upper nest, $\beta/\theta = 0.54$. They could be fitting some form of average. There is still generally good consistency between the results of GLMM and HGLM when both run to completion.

6.6 Fitting formulations similar to mixed logit by HGLM

In the previous section, HGLMs failed to fit a mixed logit formulation. This section looks for simplifications of the formulation where HGLMs do fit, although they are no longer exact mixed logits. It considers additive models, in place of multiplicative ones, and fitting regressor variables with different random distributions overall.

Additive models are closer to simple linear regression and avoid some of the approximation in GLMs; however, they can only represent rational probabilities over a restricted range. Alternative response variables can introduce less overall randomness than multinomial sampling of a logit model; the binary (0,1) result of this sampling is known to be difficult to fit with HGLMs.

6.6.1 Additive model

In the logit model, relative probabilities are formed by taking the exponential of the utility, which is commonly linear – the linear predictor of a GLM. This results in a multiplicative model. In Genstat terminolgy, there is a logarithmic link function. The formal modelling of such link functions is one of GLMs' capabilities not found in ordinary regression, but it does require iterative algorithms and some approximations.

As a simplification, the logarithmic link can be dropped (or set to 'identity'). This gives an additive model. In such a model, there is a restricted range of relative probabilities which a balancing factor can reduce to true probabilities that sum to 1.

In a multiplicative model, raw relative probabilities of 5, 3 and 2 can be reduced to true probabilities of 0.5, 0.3 and 0.2 by multiplying by a balancing factor of 0.1. In an additive model, there is no balancing factor that can be added or subtracted to give a true set of probabilities.

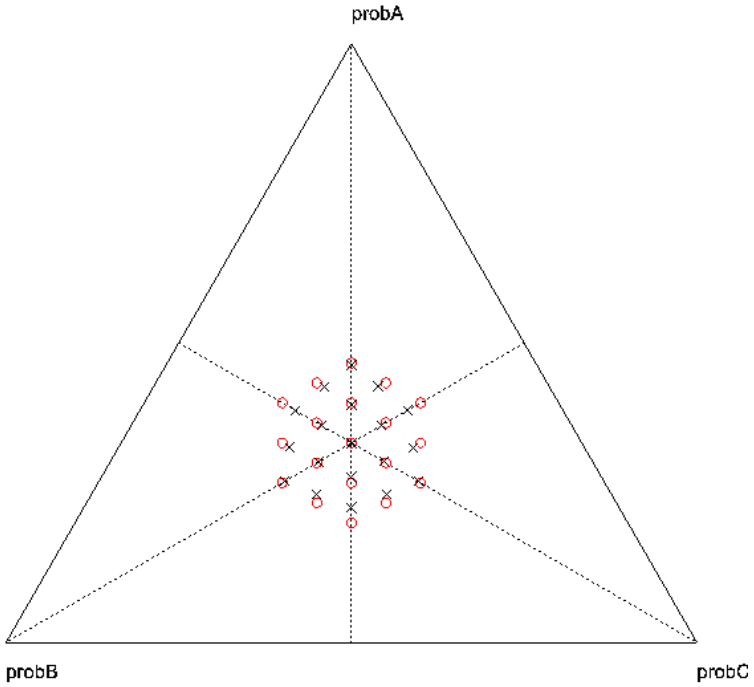
To generate data which can be balanced in an additive model, parameters of $\beta = 0.1$ and $\sigma = 0.1$ have been adopted. β complements a design matrix X of combinations of (-1,0,1) as previously. Previous tests of multiplicative datasets were based around $\beta = 1$ and $\sigma = 0.5$.

Even with the revised parameters, there is a risk of unbalanceable data being generated if extreme values are drawn for the random error component v which is scaled by σ. A whole dataset might be scaled to fit with a common scaling factor. This common factor would be incorporated in β and σ.

In an additive model, the balancing factors for individual options can shift the raw relative probabilities, but not scale their spread. If the spread is too great, something has to fall outside rational bounds. In a multiplicative model, the balancing factors can always scale the raw relative probabilities to stay within rational bounds, given that the Exponential function ensures positive values.

The red circles in figure 6.6 show the probabilities resulting from the fixed component β.X alone in an additive model. Note that they form lines parallel with the perpendicular centre-lines, whereas the lines of crosses from a multiplicative model converge on the vertices.

Figure 6.6 Systematic probabilities of additive model (β = 0.1)



o additive model

x multiplicative model, scaled to similar spread.

This additive design has to be bunched close to the centre to avoid going out of bounds with randomisation. Since multiplicative models inherently fall within these bounds, the design of the systematic component can be allowed a greater spread, as in figure 6.1 with the original value of β, 1. It seems likely this will return a better estimate of β when fitting an HGLM and may be more representative of travel demand data.

6.6.2 Model fitting

A modified version of `HGANALYSE` allowed all runs to be made with higher order Laplace transforms [`MLAPLACE=1; DLAPLACE=2`]. This had not been available for initial runs.

Previous tests had failed to recover the parameters used to generate datasets by fitting HGLMs to choice, or to Y generated with a Poisson error. Several variants had been tried.

Cases were sought where HGLMs did fit and recovered the parameters. This started with regressions on the probability μ. These can actually be fitted by ordinary regression or GLMs for a logarithmic link, if the random components Z are treated as fixed. There is then no random error in the model; the parameters are recovered almost exactly with very small residuals.

Without an overall distribution, as used to generate the Y value, probability is not strictly represented by an HGLM (or is a null case). However, as an approximation, a negligable distribution can be specified by setting `DISPERSION` small compared with other fixed and random components. This quickly converges to return the parameters.

Table 6.6 Fitting to additive normal HGLM

Overall random term Yvar			Correlating error σ^2		Systematic effect β			Iteration cycles	
Generated	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**SE fit**	**Mean**	**failed**
			0.01000	**<Expected>**	**0.10000**				
0.001	0.00100	0.00008	0.01016	0.00092	0.10035	0.00206	0.00228	6.0	
0.003	0.00301	0.00024	0.01030	0.00100	0.10052	0.00337	0.00377	7.2	
0.01	0.01004	0.00080	0.01060	0.00151	0.10063	0.00532	0.00628	8.2	
0.03	0.03013	0.00238	0.01116	0.00337	0.10062	0.00790	0.00999	12.6	
0.05	0.05017	0.00388	0.01167	0.00515	0.10051	0.00988	0.01253	21.2	1
0.1	0.09984	0.00685	0.01345	0.00839	0.10138	0.01268	0.01731	34.0	1
0.333	0.32991	0.02009	0.02340	0.02084	0.09754	0.01856	0.03089	52.6	6

SD = standard deviation

SE = standard error

Data can be generated to match this model formally, by appropriate randomisations of the probabilities to give Y variates to which HGLMs are fitted. Table 6.6 shows results for normal overall randomisation with increasing variances Yvar in the first column. This makes the Y variates continuous and not restricted to discrete values of 0 or 1 as for choice. Their expectations still sum to 1 within each case.

Results were found to vary with different random seeds, so 20 sets of randomisations were produced. The same set of 20 seeds is used for each line in the table, with different scalings by Yvar. The tabulated results are simple averages and standard deviations of parameters fitted by HGLMs to the datasets.

For small overall randomisations, all parameters are fitted well with quick convergence. As the scale of this randomisation increases, convergence slows and starts to fail (not converged after 99 iterations).

For a choice model of three options, the average probability is 1/3, and the overall variance is the same under a Poisson distribution. With a normal randomisation at this variance, at the bottom of table 6.6, HGLMs fail to converge within 99 iterations for about one in three randomisations.

The means of fitted values for Yvar and β remain close to the expected values used to generate the data. The standard deviations increase with the overall randomisation, but may still be acceptable for practical purposes. The final column for β, 'SE fit', is the simple average of standard errors output by Genstat; these are generally of the same scale and perhaps rather larger than the variation found between the 20 randomisations. (Yvar and β are antilogs of the parameters, phi and lambda Z, for which Genstat gives standard errors.)

However, the most important parameter, which HGLMs need to recover to fit a mixed logit model, is σ^2. As the overall randomness increases to 1/3, there is a serious bias in the means. Even with an overall randomness of 0.1, the standard deviation is almost as large as the expected value.

Figure 6.7 Correlating term σ^2 fitted in additive models

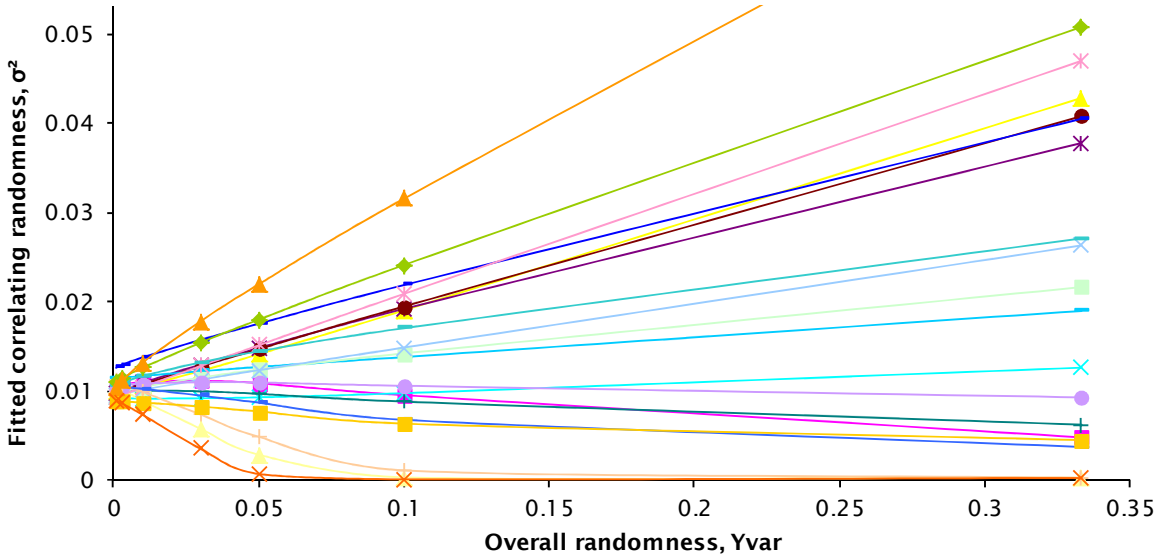


Figure 6.7 shows the values of σ^2 fitted from the 20 randomisations. Their spread appears to be scaled by the overall randomisation, and the bias when this is large may arise from a constraint to positive values.

6.6.3 Multiplicative normal

Table 6.7 Fitting to multiplicative normal HGLM

Overall random term Yvar			Correlating error σ^2		Systematic effect β			Iteration cycles	
Generated	Mean	SD	Mean	SD	Mean	SD	SE fit	Mean	Failed
			0.25000	**<Expected>**	**1.00000**				
0.001	0.00101	0.00009	0.23532	0.01908	0.99727	0.01177	0.01281	9.0	
0.003	0.00293	0.00026	0.22755	0.01881	0.98228	0.01848	0.01997	9.6	
0.01	0.00877	0.00077	0.21659	0.02397	0.93500	0.02723	0.02943	11.2	
0.03	0.02208	0.00199	0.21182	0.03376	0.85358	0.03683	0.03859	11.5	1
0.1									17

Table 6.7 shows results for multiplicative models, with logarithmic links. This is closer to the mixed logit formulation, but the overall randomisation is still normal and hence continuous. When its variance is small, HGLMs still recover the original parameters reasonably well, but there is a marked downward bias in the estimation of all parameters as the overall randomness increases. When it reaches 0.1, most models fail to fit, generally due to division-by-zero faults, possibly due to very small numbers.

Figure 6.8 Correlating term σ^2 fitted in multiplicative models

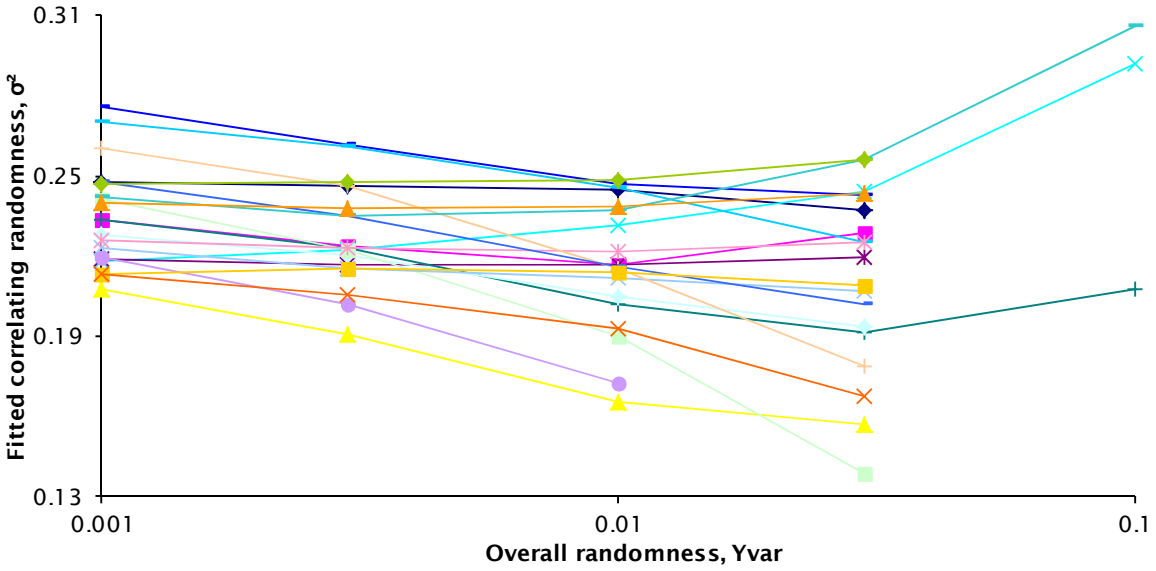


Figure 6.8 shows the general downward bias in σ^2. The trend is reversed for the three datasets which are successfully fitted for Yvar = 0.1. Note that the datasets are generated from $\beta = 1$, $\sigma = 0.5$, as in earlier work.

Additive and multiplicative normal models show different modes of failure; the log-Poisson model thought to fit a mixed logit differs again. This may mean that these findings for normal models are not indicative of the reasons for the failure to fit a mixed logit with a Poisson model; or that the reasons are multiple and extensive. It is likely that there is a lack of power in the normal datasets with large Yvar.

6.7 Simulation methods – Biogeme

Mixed logit models can be fitted by simulation methods for integration, followed by maximisation of the resulting likelihoods. These are described by Train (2003) and implemented by Bierlaire (2005) in the Biogeme package. The Biogeme package has been used as a cross-check on data generation and model fitting.

Data needed reformatting between Genstat and Biogeme. For Genstat, each option is a unit of data, with the case identified by the factor case. For Biogeme, the case is the unit of data, with information about all options included on one line. This can be seen as presenting the options in series or parallel. Genstat's `STACK` and `UNSTACK` procedures were used. Biogeme files needed trailing spaces removed, and Unix line terminators added.

Both nested and mixed models were coded and fitted.

6.7.1 Nesting

The key code sections for the nesting model were:

```
[Beta]
// Name Value    LowerBound UpperBound  status (0=variable, 1=fixed)
beta  +0.538722 -1.00e+004 +1.00e+004  0


[Utilities]
```

```
// Id Name  Avail  linear-in-parameter expression
1      A      one    beta * xA
2      B      one    beta * xB
3      C      one    beta * xC

[Model]
$NL  // Nested Logit Model
 [NLNests]
// Name paramvalue  LowerBound UpperBound  status list of alt
NESTA    +1.8562447      0         10        0      1
NESTBC   +1.8562447      0         10        0      2 3
[LinearConstraints]
NESTA - NESTBC = 0.0
```

The initial values are correct for $\beta = 1$ and $\sigma = 2$, so the code can be used to generate data with BioSim. The Biogeme model is normalised from the top, ie the parameter `beta` is specified for the top nest and then modified by θ in `NESTBC` for lower nests. Biogeme does allow normalisation from the bottom (example 10nl-bottom.mod) and section 17.10 of the manual discusses the equivalences.

A solo nest has to be specified for option A. It does not appear to matter what its parameter value is set to, but needs to be fixed. In this code, it is constrained to the same as that for the BC nest. With this constraint, `gevAlgo` has to be set to `"DONLP2"` in the default.par file.

This code appeared to retrieve parameters from the doubled dataset with $\sigma = 2$, even with initial parameters set to null values – `beta = 0, NESTA=NESTBC=1.` With the standard dataset, $\sigma = 0.5$, the retrieved `NESTBC` parameter was not significantly different from its null value, 1, indicating a lack of power.

6.7.2 Mixing

The key code sections for the mixing model were:

```
[Beta]
// Name Value  LowerBound UpperBound  status (0=variable, 1=fixed)
   A    +0.0000000e+000 -1.0000000e+000 +1.0000000e+000 1
notA    +0.0000000e+000 -1.0000000e+000 +1.0000000e+000 1
beta     1              -1.0000000e+004 +1.0000000e+004 0
sigma    2              +0.0000000e+000 +1.0000000e+004 0

[Utilities]
// Id Name  Avail  linear-in-parameter expression
1      A      one    beta * xA +    A [ sigma ] * one
2      B      one    beta * xB + notA [ sigma ] * one
3      C      one    beta * xC + notA [ sigma ] * one

[Model]
$MNL  // Multinomial Logit Model
```

Again the initial values are correct for $\beta = 1$ and $\sigma = 2$, so the code can be used to generate data with BioSim. Examples of mixed logit (logit kernel) models are given in section 18 of the Biogeme manual.

Separate error components, `A` and `notA`, have to be specified with zero mean and common `sigma`. Omitting this random term from the utility of A generates different probabilities. It would imply a heteroscedasticity which could not readily be formulated in HGLM.

This code appeared to retrieve parameters from the doubled dataset with $\sigma = 2$, giving the same log-likelihood as the nested model. However, with initial parameters set to null values (`beta=0, sigma=0`), the fitting was poor, with a lower final log-likelihood and a warning message:

```
"Unidentifiable model
 ********************
The log-likelihood is (almost) flat along the following combination of parameters ..."
```

Biogeme offers many different forms of choice model including route choice in Bioroute. Some more complex forms had to be adopted to match a basic form of HGLM. A deliberate decision was made to limit work in Biogeme in order to concentrate on the use of HGLMs in Genstat.

6.8 Random coefficients

The hypothesis that HGLMs can fit mixed logit models depends on an absence of mixing within each individual case of choice, so it has the properties of a simple logit. This property can apply to larger groups, so a coefficient may differ only between, say, car owners and non-car owners and be consistent within each group. This section examines the fit of HGLMs as the number of such groups increases from two. The overall size of the dataset is held constant, and the number of cases per group diminishes to one.

For simplicity in HGLM specification, these are random coefficient formulations, where the coefficient of utility X varies between groups, but is the same within each. If a constant (intercept rather than slope) varied between groups, but was the same within each group, it would be added into the utility of each option and cancel out in the choice. Such constants could not be recovered from a model of choice. This is a consequence of GLM's 'stacking' of data with a separate record for each option. This structure leads naturally towards generic coefficients applying to the whole of one column of data rather than alternative-specific coefficients. While alternative-specific constants are valuable in making choice models fit, generic constants are not.

The current implementation of HGLM in Genstat does not allow continuous variates such as utility X as random terms, although there is no obstacle in the theory. Another procedure, for GLMMs, is used instead. GLMMs are a subset of HGLMs, restricted to a normal distribution of random effects, and omitting recent developments in hierarchical likelihood. They employ a different algorithm.

6.8.1 Dataset

The dataset is similar to that used previously, with three options taking all combinations of (-1,0,1) for X, the sole systematic component of utility. It comprises 1080 cases, four times the standard size. This is based on a reasonable run time for GLMM. Groupings are always equally sized and as far as possible balanced in combinations; up to 40 groups, they comprise complete sets of combinations of X values ($3 \times 3 \times 3 = 27$ cases in each set).

The mean value for β the coefficient of X is again 1, but it now varies between groups about this value, normally distributed with a standard deviation σ_β set to 0.5. This raises the possibility of a negative coefficient; this might be a problem in economics, but not for statistics.

When β is randomised, the mean and variance of the generated data do not exactly match these values, particularly when there are few groups and consequently few draws of β. The generated values are shown in the 'Gen' columns of tables 6.8, 6.9 and 6.10. It can be seen that these account for a good deal of the departure of the values fitted in the models from the original values, at the head of the tables. The t values are calculated from the difference between these generated and fitted values.

Each row of the tables, with a different number and size of groups, is randomised with different seeds. These randomisations tend to obscure trends up and down the table. In particular, the draws in the top row for two groups gave close values of β, 0.882 and 1.035, and no model has been able to estimate a variance from this dataset. The same randomisation is fitted in each table.

6.8.2 Fitting by GLMM

Table 6.8 Fit by HGLM (GLMM, fixed dispersion)

Groups		Mean of β				Variance, σ_β^2			
No.	Size	Gen	Est	SE	t	Gen	Est	SE	t
Original>		1				0.25			
2	540	0.916	0.905	0.055	-0.20	0.012	0.000	bound	
5	216	0.773	0.846	0.180	0.40	0.170	0.147	0.115	-0.20
10	108	1.426	1.207	0.223	-0.98	0.262	0.450	0.236	0.80
20	54	0.977	0.890	0.095	-0.92	0.140	0.117	0.059	-0.40
40	27	0.818	0.879	0.094	0.64	0.242	0.230	0.083	-0.15
120	9	0.941	0.915	0.073	-0.36	0.245	0.254	0.080	0.11
360	3	0.904	0.873	0.059	-0.52	0.280	0.169	0.086	-1.29
1080	1	0.968	0.863	0.055	-1.89	0.263	0.041	0.115	-1.93

Table 6.8 shows results from models fitted in Genstat by:

```
GLMM [DIST=poisson; LINK=log; RANDOM=Group.X; FIXED=X+Case ] Choice
```

Estimates of the mean all fall within one standard error of the generated value, except when there is only one case per group, on the bottom line. They are also broadly reasonable estimates of the original value, 1.

Estimates of the variance are more scattered about the original value of 0.25, but in this they are generally following the generated values except at the top and bottom of the table.

The models in table 6.8 had dispersion fixed at 1. This is the default for a log-linear model, and consistent with a pure Poisson sampling process. Table 6.9 gives results from models with fitted dispersions `[DISP=*]`.

Table 6.9 Fit by HGLM (GLMM, fitted dispersion)

Groups		Mean of β				Variance, σ_β^2				Dispersion	
No.	Size	Gen	Est	SE	t	Gen	Est	SE	t	Est	SE
Original>		1				0.25					
2	540	0.916	0.905	0.055	-0.20	0.012	0.000	bound		0.990	0.030
5	216	0.773	0.846	0.180	0.40	0.170	0.147	0.116	-0.20	0.994	0.030
10	108	1.426	1.207	0.223	-0.98	0.262	0.451	0.236	0.80	0.986	0.030
20	54	0.977	0.892	0.095	-0.91	0.140	0.120	0.059	-0.35	0.964	0.030
40	27	0.818	0.883	0.095	0.68	0.242	0.241	0.085	-0.01	0.943	0.029

Groups		Mean of β				Variance, σ_β^2				Dispersion	
No.	Size	Gen	Est	SE	t	Gen	Est	SE	t	Est	SE
120	9	0.941	0.927	0.074	-0.19	0.245	0.296	0.083	0.62	0.895	0.028
360	3	0.904	0.893	0.060	-0.18	0.280	0.299	0.093	0.20	0.868	0.029
1080	1	0.968	0.915	0.057	-0.93	0.263	0.621	0.138	2.60	0.760	0.028

At the top of the table, the estimates are identical, with the fitted dispersion close to 1. This decreases with the size of group, and differences between other estimates increase. The variance is again poorly estimated when there is only one case per group, on the bottom line. However, this is now an overestimate, whereas it was an underestimate with fixed dispersion.

Convergence is generally in about five cycles. Fitting dispersion for small groups takes up to 20 cycles.

6.8.3 Fitting by simulation – Biogeme

The same datasets were analysed by the simulation package Biogeme. The key code was:

```
[PanelData]
// First, the attribute in the file containing the ID of the individual
// Then the list of random parameters which are constant for all
// observations of the same individual
// The syntax for a random paramter with mean BETA and std err SIGMA is
// BETA_SIGMA
Group_no
beta_sigma

[Beta]
// Name Value  LowerBound UpperBound  status (0=variable, 1=fixed)
beta     1     -1.00e+004 +1.00e+004  0
sigma    0.5   +0.00e+000 +1.00e+004  0

[Utilities]
// Id Name  Avail  linear-in-parameter expression (beta1*x1 + beta2*x2 + ..)
1      A      one    beta [ sigma ] * xA
2      B      one    beta [ sigma ] * xB
3      C      one    beta [ sigma ] * xC

[Model]
$MNL  // Multinomial Logit Model
```

Grouping is treated as panel data. Where there is only one case in each group, this can be omitted. Results from such models are shown at the bottom of the tables; they are the same as the models with panel data, but run more quickly.

Variance estimates and their standard errors are taken from the 'Variance of normal random coefficients' section of Biogeme reports.

Table 6.10 Fit by simulation from original initial values

Groups		Mean of β				Variance, σ_β^2				Log likelihood	Run time
No.	Size	Gen	Est	SE	t	Gen	Est	SE	t		mm:ss
Original>		1				0.25					
2	540	0.916	0.905	0.055	-0.20	0.012	4.2E-11	9.9E-07	####	-1024.2	02:52
5	216	0.773	0.818	0.136	0.33	0.170	0.126	0.081	-0.54	-1039.1	01:48
10	108	1.426	1.159	0.189	-1.42	0.262	0.472	0.218	0.96	-940.7	01:54
20	54	0.977	0.915	0.095	-0.65	0.140	0.117	0.062	-0.39	-1029.6	01:56
40	27	0.818	0.916	0.099	0.99	0.242	0.239	0.092	-0.04	-1032.5	01:24
120	9	0.941	0.984	0.082	0.52	0.245	0.318	0.104	0.71	-1021.6	00:46
360	3	0.904	0.954	0.075	0.66	0.280	0.261	0.127	-0.15	-1034.3	00:34
1080	1	0.968	0.895	0.109	-0.66	0.263	0.089	0.268	-0.65	-1036.6	01:03
No panel		0.968	0.895	0.109	-0.66	0.263	0.089	0.268	-0.65	-1036.6	00:37

Table 6.10 shows a similar pattern of results to table 6.8. However, they match only in the mean for two groups where variances have taken very low values. Even where estimates of variance coincide for 20 groups, the estimates of the means do not.

Again, estimates for the variance are poor where there is one case per group, but here it is accommodated in a much large standard error.

The models in table 6.10 were given the original values as starting points, which cannot be set for GLMM (except perhaps in the REML procedures from which it is built). Starting from null values:

```
[Beta]
// Name Value  LowerBound UpperBound  status (0=variable, 1=fixed)
beta     0     -1.00e+004 +1.00e+004  0
sigma    0     +0.00e+000 +1.00e+004  0
```

Biogeme gives the results shown in table 6.11.

Table 6.11 Fit by simulation from null initial values

Groups		Mean of β				Variance, σ_β^2				Log likelihood	Run time
No.	Size	Gen	Est	SE	t	Gen	Est	SE	t		mm:ss
Original>		1				0.25					
2	540	0.916	0.905	0.055	-0.20	0.012	2.2E-16	1.0E-16	####	-1024.2	11:25
5	216	0.773	0.818	0.136	0.33	0.170	0.126	0.081	-0.54	-1039.1	07:12
10	108	1.426	1.055	0.059	-6.28	0.262	2.2E-16	*	*	-979.2	16:59
20	54	0.977	0.863	0.055	-2.10	0.140	2.2E-16	1.3E-16	####	-1036.6	04:05
40	27	0.818	0.916	0.099	0.99	0.242	0.239	0.092	-0.04	-1032.5	03:16
120	9	0.941	0.875	0.055	-1.21	0.245	1.2E-08	*	*	-1033.1	02:37
360	3	0.904	0.860	0.054	-0.81	0.280	2.2E-16	3.9E-16	####	-1037.5	01:01
1080	1	0.968	0.863	0.055	-1.92	0.263	2.9E-07	*	*	-1036.6	00:48
No panel		0.968	0.863	0.055	-1.92	0.263	2.9E-07	*	*	-1036.6	00:28

For most groupings the estimate of variance takes an extremely low value, as has been seen in previous unsuccessful fittings of HGLMs to mixed models. For these groupings the estimates of the mean differ from table 6.10 and the log-likelihoods are higher, except where there is one case per group (bottom rows) where the log-likelihood is the same.

Where the estimate of variance is not small, for 5 and 40 groups, the results match those in table 6.10.

Run times are generally longer and more varied than for table 6.10.

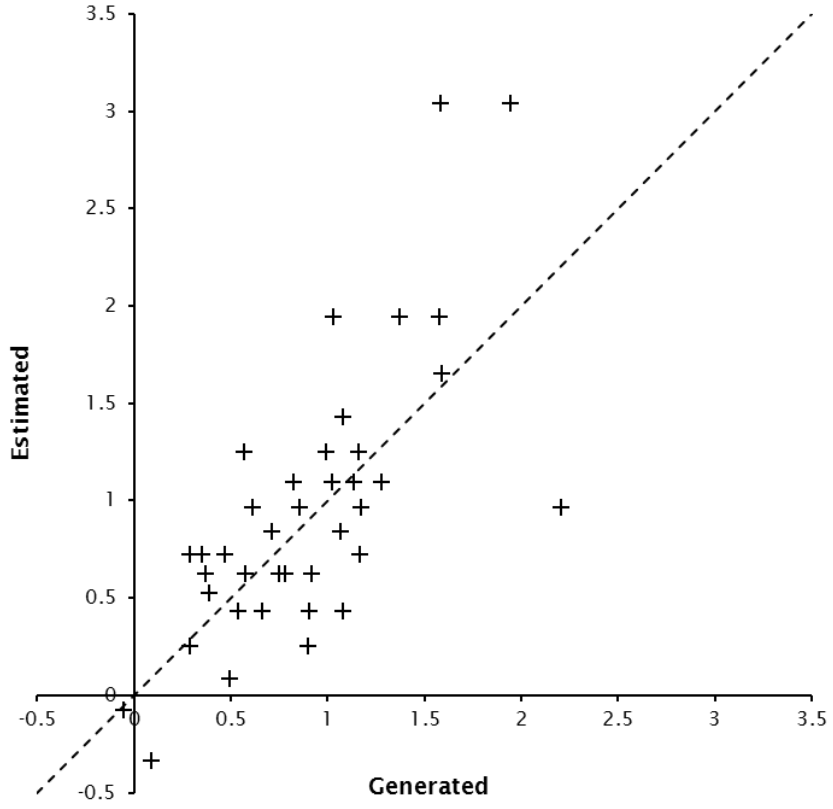
6.8.4 Fixed effects – GLM

The coefficient for each group can be fitted by as an ordinary log-linear GLM, treating it as a fixed effect.

```
MODEL [DIST=poisson; LINK=log; GROUP=Case] Choice
FIT Group.X
```

The coefficients thus estimated for 40 groups are compared with the values used in generating the dataset in figure 6.9.

Figure 6.9 Coefficients β of 40 groups



Individual regressions for each group give the same estimates as a single analysis of the whole dataset, except for one extreme value (group 16, outlying figure 6.9: 1.54 generated, 11 estimated in separate model, 6.48 estimated in joint model – perhaps iteration finished sooner when other coefficients fitted well).

Regressions on the probabilities underlying the multinomial sampling of a single choice returned the generated values exactly.

6.9 Error components

A random coefficient formulation was more successful where the groups between which the coefficient varied were larger than the individual case. This section applies these groupings to the original error component formulation.

The datasets are the same size as for the random coefficient tests, 1080 cases. This is four times the size of the standard dataset and twice the size of the larger dataset adopted after considering power of detection. The scale of the random component σ is taken as 2 as in that later work, rather than the original standard value of 0.5. As in all previous work, there are three options/alternatives, taking all combinations of {-1,0,1} for utility X. Its coefficient β is fixed at 1 again, after varying about that value in the random coefficient formulation.

For an error component formulation, HGLM can be used as well as GLMM. The random term distribution is taken as normal for compatibility with GLMM.

The mean and variance of random terms in generated datasets can differ considerably from those of their parent distributions, making the fit of estimates from models hard to assess. For this analysis, the datasets for each size of group were not randomised independently. A single randomisation was drawn for the maximum number of groups, one group for every case, the bottom line of the tables. The draws were averaged for larger groupings and rescaled to the original variance. This process may have brought a little more consistency between the lines in the tables, but the main effects of randomisation remain and a lack of independence between lines hampers interpretation. To meet this, sets of three tables have been produced with independent randomisations, with seed = 0, 1 or 2.

The 'Gen' column of the tables shows that the variance of the generated random components is below the original value of 4 for all groupings of seed = 0, but particularly for small numbers of groups at the top of the table. This effect is also seen for small number of groups with seed = 2, but the generations from seed = 1 fall more equally around the original value of 4.

t statistics are calculated for the difference of the estimate from the generated value, or from the original value 1 for β. t0 is the estimate's difference from 0.

The bottom line of each table, with one group per case, is the formulation studied in the earlier part of this chapter.

6.9.1 GLMM

GLMMs are fitted by Genstat code:

```
GLMM [DIST=poisson; LINK=log; RANDOM=Group.A; FIXED=Case+X] Choice
```

where A is a dummy variable or factor for option A, rather than B or C. Dispersion is fixed at 1 by default.

Table 6.12 Fit by GLMM algorithm

Groups		β			σ²					Cycles
No.	Size	Est	SE	t	Gen	Est	SE	t	t0	
Original>		1			4					
Seed = 0										
2	540	0.888	0.066	-1.70	1.601	1.979	1.987	0.19	1.00	8
5	216	0.844	0.063	-2.48	1.694	2.008	1.288	0.24	1.56	11
10	108	0.905	0.062	-1.55	2.016	1.875	0.900	-0.16	2.08	11
20	54	0.879	0.063	-1.92	2.401	2.327	0.787	-0.09	2.96	8
40	27	0.860	0.062	-2.27	2.842	1.968	0.493	-1.77	3.99	11
120	9	0.861	0.070	-2.00	3.557	1.997	0.336	-4.64	5.94	8
360	3	0.756	0.064	-3.80	3.780	1.533	0.198	-11.35	7.74	11
1080	1	0.629	0.052	-7.11	3.925	0.380	0.118	-30.04	3.22	7
Seed = 1										
2	540	0.993	0.068	-0.11	4.452	5.435	5.476	0.18	0.99	13
5	216	1.024	0.064	0.37	3.058	2.980	1.943	-0.04	1.53	10
10	108	1.054	0.065	0.84	3.249	2.670	1.256	-0.46	2.13	8
20	54	1.013	0.065	0.20	3.385	2.833	0.987	-0.56	2.87	12
40	27	0.959	0.064	-0.63	4.292	2.927	0.747	-1.83	3.92	11
120	9	1.039	0.074	0.53	3.807	2.552	0.423	-2.97	6.03	10
360	3	0.848	0.065	-2.36	4.168	1.450	0.193	-14.08	7.51	11
1080	1	0.713	0.053	-5.38	4.220	0.338	0.118	-32.90	2.86	6
Seed = 2										
2	540	0.887	0.058	-1.96	1.117	0.965	0.977	-0.16	0.99	5
5	216	0.958	0.060	-0.70	2.022	0.941	0.610	-1.77	1.54	6
10	108	1.010	0.067	0.15	3.268	3.622	1.726	0.20	2.10	10
20	54	0.976	0.067	-0.36	4.436	3.821	1.318	-0.47	2.90	15
40	27	0.965	0.068	-0.52	4.532	3.877	0.977	-0.67	3.97	13
120	9	0.885	0.072	-1.59	3.985	2.590	0.428	-3.26	6.05	11
360	3	0.769	0.063	-3.65	4.092	1.317	0.179	-15.50	7.36	10
1080	1	0.643	0.052	-6.84	4.076	0.335	0.116	-32.25	2.89	6

The coefficient of utility β is generally underestimated for seed = 0, but particularly for many small groups, at the bottom of the table. This underestimation for small groups is also apparent with the other two seeds where other estimates appear reasonable.

Estimates of σ^2 also appear reasonable for the upper part of the tables once the variability in the generated values is taken into account. However, serious underestimation appears in the bottom three lines of the table. This underestimation of the correlating effect of the random component may lead to the estimate for β reflecting the upper nest coefficient of 0.54, as well as the lower nest coefficient of 1.

Except for small numbers of groups, at the top of the tables, t0 gives confidence that there is a random error component.

6.9.2 HGLM

HGLMs are fitted by Genstat code:

```
HGRANDOM [DIST=normal; LINK=identity] Group.A
HGFIXED [DIST=poisson; LINK=log] Case,X
HGANALYSE [MLAPLACE=1; DLAPLACE=2] Choice
```

This specifies higher-order Laplace approximations, using modified code for `HGANALYSE`, but otherwise defaults are accepted, including dispersion fixed at 1.

The bottom two groupings, with many levels in the random term, fail for lack of data storage space on the computer. This was encountered when choosing a more powerful dataset for the main study and led to a sample size of 540 cases rather than 1080, so that the formulation of one group per case in the bottom line could be fitted within the constraints of space.

Table 6.13 Fit by HGLM algorithm

Groups		β			σ^2					Cycles
No.	Size	Est	SE	t	Gen	Est	SE	t	t0	
Original>		1			4					
Seed = 0							*Log scale*			
2	540	0.888	0.066	-1.69	1.601	1.991	*0.999*	*0.22*		2
5	216	0.844	0.063	-2.48	1.694	1.999	*0.633*	*0.26*		2
10	108	0.905	0.061	-1.57	2.016	1.837	*0.450*	*-0.21*		5
20	54	0.878	0.063	-1.94	2.401	2.283	*0.319*	*-0.16*		5
40	27	0.857	0.061	-2.35	2.842	1.853	*0.226*	*-1.89*		5
120	9	0.850	0.069	-2.17	3.557	1.519	*0.132*	*-6.45*		8
360	3	Out of space								
1080	1									
Seed = 1							*Log scale*			
2	540	0.993	0.068	-0.10	4.452	5.561	*1.010*	*0.22*		4
5	216	1.024	0.064	0.37	3.058	2.955	*0.636*	*-0.05*		4
10	108	1.054	0.064	0.84	3.249	2.633	*0.451*	*-0.47*		4
20	54	1.011	0.065	0.17	3.385	2.699	*0.317*	*-0.71*		5
40	27	0.955	0.064	-0.70	4.292	2.774	*0.227*	*-1.92*		8
120	9	1.026	0.073	0.35	3.807	2.153	*0.134*	*-4.25*		12
360	3	Out of space								
1080	1									
Seed = 2							*Log scale*			
2	540	0.887	0.057	-1.98	1.117	0.959	*1.000*	*-0.15*		4
5	216	0.958	0.060	-0.70	2.022	0.926	*0.634*	*-1.23*		4
10	108	1.009	0.067	0.14	3.268	3.523	*0.447*	*0.17*		6
20	54	Failed to converge								99
40	27	0.958	0.067	-0.63	4.532	3.490	*0.221*	*-1.18*		5
120	9	0.877	0.071	-1.73	3.985	2.163	*0.133*	*-4.59*		8
360	3	Out of space								
1080	1									

Estimates for β are very close to those from GLMMs at the top of the tables, but show a slightly greater tendency to underestimate lower down the table.

Estimates for σ^2 also correspond with those from GLMMs at the top of the tables, but show an even greater tendency to underestimation lower down the table.

Standard errors are given on the log scale, as output in Genstat, and the t statistics are calculated on that scale. It is not possible to calculate a t0 statistic thus, because log(0) is not defined.

6.9.3 Biogeme

The key code for Biogeme, initialised with original values, is:

```
[PanelData]
// First, the attribute in the file containing the ID of the individual
// Then the list of random parameters which are constant for all
// observations of the same individual
// The syntax for a random paramter with mean BETA and std err SIGMA is
// BETA_SIGMA
Group_no
A_sigma
notA_sigma

[Beta]
// Name Value  LowerBound UpperBound  status (0=variable, 1=fixed)
   A    +0.0000000e+000 -1.0000000e+000 +1.0000000e+000 1
notA    +0.0000000e+000 -1.0000000e+000 +1.0000000e+000 1
beta     1              -1.0000000e+004 +1.0000000e+004 0
sigma    2              +0.0000000e+000 +1.0000000e+004 0

[Utilities]
// Id Name  Avail  linear-in-parameter expression (beta1*x1 + beta2*x2 + ..)
1      A      one    beta * xA +    A [ sigma ] * one
2      B      one    beta * xB + notA [ sigma ] * one
3      C      one    beta * xC + notA [ sigma ] * one

[Model]
$MNL  // Multinomial Logit Model
```

Estimates for the variance of the error component σ^2 are taken from the 'utility parameters' section of Biogeme's reports. The 'Variance of normal random coefficients' output, quoted in random parameter analysis, gives two values for `A_sigma` and `notA_sigma`. These usually correspond with each other, but not always with the square of the value of `sigma` in 'utility parameters', which appears more rational. The square of `sigma` is shown in the tables, but its standard error and the t statistics derived from it are on the linear scale.

Table 6.14 Fit by Biogeme from original initial values

Groups		β			σ²					Log likelihood	Run time
No.	Size	Est	SE	t	Gen	Est	SE	t	t0		
Original>		1			4						mm:ss
Seed = 0							*Linear scale*				
2	540	0.890	0.066	-1.68	1.601	2.947	*0.191*	*2.36*	*8.98*	-771.3	11:09
5	216	0.846	0.063	-2.44	1.694	3.298	*0.259*	*1.99*	*7.01*	-813.1	11:40
10	108	0.910	0.062	-1.45	2.016	2.148	*0.406*	*0.11*	*3.61*	-868.8	4:13
20	54	0.888	0.064	-1.74	2.401	2.505	*0.280*	*0.12*	*5.64*	-826.1	2:03
40	27	0.879	0.063	-1.92	2.842	2.309	*0.206*	*-0.81*	*7.38*	-866.4	1:19
120	9	0.910	0.074	-1.22	3.557	3.394	*0.192*	*-0.23*	*9.58*	-885.7	0:54
360	3	0.908	0.077	-1.21	3.780	4.439	*0.190*	*0.86*	*11.11*	-955.9	0:39
1080	1	0.915	0.083	-1.02	3.925	4.557	*0.366*	*0.42*	*5.83*	-1079.5	0:54
No panel		0.915	0.083	-1.02	3.925	4.557	*0.366*	*0.42*	*5.83*	-1079.5	0:31
Seed = 1							*Linear scale*				
2	540	0.993	0.068	-0.10	4.452	3.541	*0.514*	*-0.44*	*3.66*	-721.8	1:56
5	216	1.026	0.064	0.40	3.058	3.112	*0.504*	*0.03*	*3.50*	-832.2	2:58
10	108	1.059	0.065	0.90	3.249	2.634	*0.419*	*-0.43*	*3.87*	-839.7	1:25
20	54	1.023	0.066	0.34	3.385	3.577	*0.471*	*0.11*	*4.01*	-827.7	1:25
40	27	0.981	0.066	-0.30	4.292	3.906	*0.271*	*-0.35*	*7.28*	-827.3	2:58
120	9	1.107	0.080	1.35	3.807	4.521	*0.226*	*0.78*	*9.42*	-835.9	0:53
360	3	1.012	0.078	0.16	4.168	4.334	*0.195*	*0.21*	*10.69*	-947.9	0:49
1080	1	0.968	0.084	-0.38	4.220	2.956	*0.291*	*-1.15*	*5.91*	-1063.5	0:41
No panel		0.968	0.084	-0.38	4.220	2.956	*0.291*	*-1.15*	*5.91*	-1063.5	0:24
Seed = 2							*Linear scale*				
2	540	0.887	0.058	-1.95	1.117	0.981	*0.303*	*-0.22*	*3.26*	-945.7	47:35
5	216	0.961	0.060	-0.65	2.022	0.950	*0.135*	*-3.31*	*7.21*	-916.8	3:21
10	108	1.017	0.068	0.24	3.268	3.251	*0.340*	*-0.01*	*5.31*	-774.7	3:15
20	54	0.987	0.068	-0.20	4.436	4.568	*0.385*	*0.08*	*5.54*	-773.4	1:10
40	27	0.988	0.070	-0.18	4.532	5.433	*0.364*	*0.56*	*6.41*	-762.9	0:53
120	9	0.939	0.077	-0.79	3.985	4.832	*0.233*	*0.87*	*9.42*	-835.9	0:50
360	3	0.906	0.075	-1.25	4.092	3.597	*0.176*	*-0.72*	*10.78*	-970.9	0:31
1080	1	0.890	0.082	-1.34	4.076	3.182	*0.314*	*-0.75*	*5.68*	-1081.7	0:41
No panel		0.890	0.082	-1.34	4.076	3.182	*0.314*	*-0.75*	*5.68*	-1081.7	0:24

As with regression methods, there is a general underestimation of β with seed = 0. At the top of all three tables, estimates of β are very similar to those from regression, but differences increase down the tables, and there is no clear underestimation for small groups at the bottom of the tables.

There is little correspondence with the regression methods in the estimates for σ^2, except perhaps at the top of the table for seed = 2. There is a reasonable correspondence with the generated values throughout the tables.

The t0 statistic gives confidence that there is a random error component throughout the tables. The few, larger groups at the top of the table require more run time.

When Biogeme is initialised with null values:

```
[Beta]
// Name Value  LowerBound UpperBound  status (0=variable, 1=fixed)
beta     0              -1.0000000e+004 +1.0000000e+004 0
sigma    0              +0.0000000e+000 +1.0000000e+004 0
```

it often fits extreme values for σ^2.

Table 6.15 Fit by Biogeme from null initial values

Groups		β			σ²					Log likelihood	Run time
No.	Size	Est	SE	t	Gen	Est	SE	t	t0		
Original>		1			4						mm:ss
Seed = 0							*Linear scale*				
2	540	0.890	0.066	-1.67	1.601	0.432	*0.043*	*-14.26*	*15.44*	-773.1	55:37
5	216	0.847	0.063	-2.43	1.694	1.238	*0.283*	*-0.67*	*3.94*	-813.7	30:52
10	108	0.911	0.062	-1.45	2.016	2.148	*0.406*	*0.11*	*3.61*	-868.8	13:02
20	54	0.888	0.064	-1.74	2.401	2.505	*0.280*	*0.12*	*5.64*	-826.1	7:43
40	27	0.638	0.050	-7.17	2.842	2.2E-16	*1.8E308*	*	*	-1097.9	1:16
120	9	0.669	0.051	-6.50	3.557	2.2E-16	*2.3E-09*	*	*	-1090.1	3:02
360	3	0.601	0.050	-8.01	3.780	2.2E-16	*2.9E-09*	*	*	-1107.2	1:27
1080	1	0.611	0.050	-7.79	3.925	2.2E-16	*1.8E308*	*	*	-1104.8	0:43
No panel		0.611	0.050	-7.79	3.925	2.2E-16	*1.8E308*	*	*	-1104.8	0:23
Seed = 1							*Linear scale*				
2	540	0.701	0.051	-5.82	4.452	2.2E-16	*1.0E-09*	*	*	-1081.9	8:59
5	216	1.026	0.064	0.40	3.058	3.111	*0.505*	*0.03*	*3.50*	-832.2	17:56
10	108	1.059	0.065	0.90	3.249	2.634	*0.419*	*-0.43*	*3.87*	-839.7	9:07
20	54	1.023	0.066	0.34	3.385	3.574	*0.471*	*0.11*	*4.01*	-827.7	6:49
40	27	0.981	0.066	-0.30	4.292	3.906	*0.271*	*-0.35*	*7.28*	-827.3	5:01
120	9	1.107	0.080	1.35	3.807	4.521	*0.226*	*0.78*	*9.42*	-835.9	1:41
360	3	1.012	0.078	0.16	4.168	4.334	*0.195*	*0.21*	*10.69*	-947.9	1:43
1080	1	0.701	0.051	-5.82	4.220	1.1E-08	*1.8E308*	*	*	-1081.9	0:53
No panel		0.701	0.051	-5.82	4.220	1.1E-08	*1.8E308*	*	*	-1081.9	0:30
Seed = 2							*Linear scale*				
2	540	0.887	0.058	-1.95	1.117	0.979	*0.305*	*-0.22*	*3.24*	-945.7	65:44
5	216	0.828	0.054	-3.20	2.022	2.2E-16	*6.2E-09*	*	*	-1046.8	7:15
10	108	0.706	0.052	-5.71	3.268	2.2E-16	*1.8E308*	*	*	-1080.5	2:26
20	54	0.669	0.051	-6.50	4.436	2.2E-16	*1.8E308*	*	*	-1090.1	3:51
40	27	0.988	0.070	-0.18	4.532	5.433	*0.364*	*0.56*	*6.41*	-762.9	3:54
120	9	0.646	0.051	-7.00	3.985	2.2E-16	*1.8E308*	*	*	-1096.0	4:03
360	3	0.659	0.051	-6.72	4.092	2.2E-16	*4.1E-09*	*	*	-1092.7	1:27
1080	1	0.628	0.050	-7.40	4.076	2.2E-16	*1.8E308*	*	*	-1100.5	0:43
No panel		0.628	0.050	-7.40	4.076	2.2E-16	*8.5E-09*	*	*	-1100.5	0:24

The only consistent fitting of non-extreme values is in the middle of the table for seed = 1, where the randomisations are closer to the original values from which they are drawn. These results agree with those of initialised models, but the results in the top two lines for seed = 0 do not. Log-likelihoods for these and all extreme-value solutions are higher than for solutions from initialised models.

6.10 Further approaches and issues

This section discusses further approaches to the problem, some of which have already been pursued to a limited extent, and areas where the hypothesis might be at fault. In combination with the variants already tried, there are so many possibilities that the next step needs to be a better fundamental understanding of HGLMs as suggested in the next two sections and of existing methods for choice modelling in the last section.

6.10.1 Likelihood surfaces

Nelder (pers comm) commented that the poor convergence found when trying to fit HGLMs was often indicative of a flat likelihood surface without a single distinct maximum. Plotting curves and surfaces of probabilities and likelihoods was started for a single option for a single case, extending to all options for a single case. Even this generated a large dataset with for all possible combinations of β, σ, and the random components for A and BC.

This did not give any immediate insight. Further work was needed to aggregate over all cases in a dataset and perhaps over possible outcomes. It still seemed to offer a visualisation of the processes at the heart of HGLMs.

A dichotomous point distribution for the random coefficient $v = \pm\sigma$ would much simplify the enumeration of possible values compared with continuous distributions such as the normal, gamma and beta (tried in a spreadsheet).

6.10.2 Own calculation of HGLM

Another approach to understanding HGLMs was to replicate calculations from Lee et al (2006) for the trial dataset. Rather than starting from scratch, all data structures were retrieved from the Genstat workspace 'G5PL_HG', where the workings and products from fitting an HGLM are stored. Lee et al (2006) equation 5.24 for likelihood was calculated in this way.

This retrieval of Genstat workings also facilitates a closer examination of other models, such as the examples given in Lee et al (2006).

6.10.3 Fixing known values

Since the trial datasets are generated entirely from known values, there are opportunities to use these 'true' rather than estimated values in different parts of the model fitting process. Systematic effects might be fixed as part of the OFFSET in Genstat GLMs; these could include the realisations of random components. Dispersions might be fixed as WEIGHT; where this allows a matrix, correlations could also be included.

These methods might be used to estimate one of the unknown parameter at a time, and contrast the workings and results with the usual case when all parameters are unknown.

6.10.4 Aggregation within groups

Where groups with common randomisations or tastes have been generated for random coefficients or error components, the dataset might be compacted. Cases with common X values could be aggregated,

with the {0,1} choice indicator summed to count the number of times each option is chosen. If there were, say, 10 repeated cases, they could be summarised for HGLM as just three records, eg:

A Xa=0 chosen 3 times

B Xb=1 chosen 5 times

C Xc=-1 chosen 2 times

instead of 3x10=30 records as individual cases. Biogeme's data format could be similarly compacted. This might relieve computational problems even if the statistical problem is exactly equivalent.

Aggregation across different X values would lose information in averaging them.

6.10.5 Redundancy and aliasing between balancing factors and random terms

For each case, there are three items of data, describing the three options and their outcomes. There are also three parameters to be fitted:

- balance, the coefficient of the case factor
- v_A, the random coefficient for the level of Z for option A
- v_{BC}, the random coefficient for the level of Z for options B and C.

These are not completely redundant in fitting the outcomes of the options completely, because all parameters are common to B and C, and no combination of the parameters can fit a difference in the outcomes for B and C.

For the outcome choice to be the same for B and C they must both be zero; they cannot both be one. This will give an 'empty nest' at the bottom of a nesting model. If v_{BC} were a fixed effect like balance it could provide a good fit just by taking a very low value, providing no information about β but diluting the measure of fit like an 'empty zone'.

The parameters are aliased; Genstat recognises this when fitting both case and Z as fixed terms in an ordinary GLM. It drops some levels and fits extreme values for many others.

The only further constraint when Z is entered as a random term in HGLM or GLMM seems to be that its coefficients v should be normally distributed ~N(0, σ^2). Is the condition of a normal shape sufficient to prevent a very large σ being estimated for a distribution with a lot of extreme values and nothing in the middle?

Such a 'hole in the middle' distribution could be deliberately generated, eg the dichotomous random distribution suggested for curve/surface plotting in section 6.11.1. Can HGLM tell the difference between a doughnut and a bun for its random distribution?

HGLMs can fit random error terms that vary between individual units ('bottom level'), which are open to aliasing with any other term in the model.

6.10.6 Correlation between balancing factors and random terms.

If aliasing above is not complete, there may still be correlation between the balance and v – such as in repeated measurement analysis (Lee et al 2006, p138). Train (2003) defines μ as having zero mean in section 6.3 on error components, but considers random coefficients β with non-zero means in section 6.2 for another interpretation of mixed logit models. Genstat recognises both interpretations for fitting by REML, a normal subset of HGLMs. Lee et al (2006, section 6.1.1) discusses arbitrary constraints on location, leading to differences between HGLMs and GLMMs.

In generating the datasets, balancing factors are calculated using the random terms after they have been drawn from randomisations, so the balancing factors incorporate some information about the random terms.

Balancing factors have similar correlations with other systematic terms, but this does not appear to affect the fitting of a plain multinomial logit model by GLM. Possibly the correlations are less critical when the systematic coefficient β is estimated over the whole dataset, but individual v_A and v_{BC} are fitted within each case.

6.10.7 Correlation between choice of options

When a dataset is generated with an independent Poisson distribution for each option, as would be expected from the HGLM specification, the resultant Y values are not constrained to sum to 1 for each case. This constraint does apply to choice, the dependent variable for a logit choice model.

The balancing factor scales the probabilities so that the expectations of Y sum to 1. In doing so, it accounts for one degree of freedom due to the constraint. However, there are still correlations in Choice that are not present in Y.

There is asymmetry in the pattern of correlation. If $\text{Choice}_A = 0$, then the probability that $\text{Choice}_B = 1$ increases, say from 33% to 50% if underlying probabilities are equal. However, if $\text{Choice}_A = 1$, Choice_B has to be 0.

Empirically, this correlation has not been seen to affect plain multinomial logit models fitted by GLMs. It could be interacting with the mixing element of the logit model or the random component in the HGLM. However, if this were the case it could be expected HGLMs would fit to Y but not to choice, and this has not been observed in testing even with a dataset of adequate power.

In the limiting case of two options, the correlation between choices becomes entirely systematic, $\text{choice}_B = 1 - \text{choice}_A$. This allows a GLM to be fitted to just one item of data per case, representing differences between options. Since there is no longer a balancing factor, the number of degrees of freedom remains consistent. The fixed/sampling distribution becomes Binomial with a logit link.

6.10.8 Binomial fixed/sampling distribution

Even with three or more options, the variable choice can only take binary values 0 and 1. The Poisson distribution allows the possibility of all other positive integers as well. This suggests the binomial distribution may be more appropriate. A logit rather than logarithmic link is most common with a binomial distribution. A quick test of these options gave no obviously promising results.

6.10.9 Gamma random/mixing distribution

In the standard trial dataset, the random component common to options B and C has a normal distribution for simplicity and backward compatibility with other models such as the GLMM. This is not conjugate with the Poisson distribution of overall sampling, which makes the fitting of an HGLM more complex.

The gamma distribution is conjugate with the Poisson, so might help the fitting of an HGLM.

A Poisson distribution about a gamma-distributed mean gives a negative binomial distribution. However, in the trial scenario, options B and C will have separate Poisson distributions about a common gamma mean.

This approach was not tried because of uncertainty as to how to generate a gamma distribution with given variance and logarithmic link for the random term, or what parameter would be recovered from an HGLM.

6.10.10 Small-variance Poisson distributions

Section 6.6 showed HGLMs can recover parameters from datasets where the overall randomness is relatively small and normally distributed. The true Poisson distribution implies a relatively large randomness, but there are under-distributed variants of the Poisson such as the quasi-Poisson. These may help distinguish the effects of the scale of the Poisson's variance from some of its other properties. However, they would depart from the multinomial sampling process implicit in generating mixed logit data.

The distinction between discrete and continuous distributions may be more important than the relationship between variance and mean, particularly if the discrete distribution is restricted to 0 and 1 as for choice.

Fitting under-dispersed distributions could involve or explore the effects of the `DISPERSION` option of `HGFIXEDMODEL`.

6.10.11 Alternative methods within HGLM

HGLMs are an extensive superset of modelling methods that include ordinary regression and analysis of variance, GLMs, REML models, GLMMs and general estimating equations (GEE). Many of these are applicable to particular cases of mixed logit or to similar formulations, particularly where distributions are normal. They also offer alternative algorithms which may be contrasted with HGLMs or for parts of the process as suggested in section 6.11.3.

This approach has already been taken in fitting GLMMs in section 6.5 and introducing additive formulations and normal overall distributions in section 6.6.

Other statistical software than Genstat may offer further alternative methods and algorithms.

6.10.12 Alternative methods for mixed logit

All but the simplest choice models are usually fitted using specialist software such as Alogit, Nlogit or Biogeme. Biogeme has been used for some cross-checking in section 6.7. The package offered many avenues for exploration, but work with it was deliberately limited to concentrate on the HGLM approach.

According to Prof. John Polak (Imperial College, pers comm), while the simple multinomial logit model has a convex likelihood, other forms of discrete choice model, including mixed logit, do not in general; considerable work has gone into algorithms to find the global maximum in other forms. The theory and practice that has been developed would be a thorough grounding in the topic.

6.11 Summary

6.11.1 Theoretical

No contradictions to the hypothesis that HGLMs can fit mixed logit models have been recognised. A number of possibilities are suggested in section 6.10.

6.11.2 Empirical

Empirical proofs have failed, including a number of closely related variants. The initial datasets probably lacked power, but some repetitions with more powerful datasets gave no better results.

GLMs and HGLMs can recover parameters from probabilities similar to mixed logit formulations in that they comprise fixed, random and balancing factors. Therefore there is no fundamental and universal barrier to HGLMs fitting such a formulation, with its close relationships between random and balancing factors.

HGLMs can fit some randomisations of the probabilities, but if the variance is large the fitting fails or is biased. The all-or-nothing choice between one alternative or another adds a large amount of randomness to the underlying probability, obscuring the mixing effects.

HGLMs give better results for fewer, larger groups with coefficients or error components in common, only differing between groups. Disaggregate models typically have a different randomisation for every case.

GLMMs generally gave similar results to HGLMs. If the two algorithms are independent, this suggests that problems lie outside the algorithms.

Simulation methods can also encounter difficulties fitting mixed logit models to this dataset; the equivalent nested models may be better conditioned.

Both regression and simulation methods offer many options for improving their performance. Biogeme offers different optimisers. A search for better performance needs to be directed by a better knowledge of the algorithms for fitting mixed logit models and of their underlying properties.

6.11.3 Efficiency

Existing methods of fitting mixed logit models require simulation methods to calculate integrals. One potential benefit of HGLMs is that they do not do this and hence may be more computationally efficient. Experience to date is that HGLMs take longer to run than the simulation method Biogeme.

The computation of HGLMs is understood to expand considerably with the number of cases. The number of levels of both the balancing factor and the random terms increase accordingly, and with them the size of matrices being manipulated. The number of cases could be large in practical application. HGLMs may well not give efficient solutions unless absorption/grouping can be incorporated. This has been introduced for random effects in later versions of Genstat, but the balancing factor remains a fixed effect with one level for each case.

7 Spatial patterns

7.1 Introduction

This chapter investigates spatial patterns in trip distribution which go beyond the cost deterrence effects of simple models. It treats these patterns as random errors which can be estimated by hierarchical general linear models (HGLMs) according to geospatial theory. However, HGLMs proved difficult to fit on the scale of the Wellington model and several associated spatial aspects have been explored.

Although sampling errors can account for errors in fitting trip distribution models, their fit to screenline counts was not particularly good, as has been observed in many practical models. This suggests that individual errors are small compared with those of sampling, and would tend to cancel out if independent. For errors to appear substantial there needs to be a correlating factor and this could be spatial.

A conventional response to accommodate spatial variations in trip distributions is the introduction of K factors (see section 1.3.6). K factors represent blocks of movements with all-or-nothing correlations. Continuous variation in spatial correlation can also be modelled in HGLMs. This gives two approaches to spatial correlation:

Top down: Fitting K factors for progressively finer segmentations.

Bottom up: Looking for an improvement in fit from a correlation between zone-to-zone movements that is a function of their separation.

Both these approaches encountered computational problems, particularly limits on computer space. These were addressed in part by grouping zones into sectors, which is intrinsic to applying K factors.

K factors are conventionally fitted as part of the systematic model. They can take any value to match modelled trips with observations in their segment and will achieve an exact match when fitted by a log-linear Poisson GLM. As part of the random model considered in this chapter, the K factors are also expected to conform to a probability distribution, which tends to limit the more extreme values.

The application of random K factors can be seen as a form of hierarchical model similar to the nested logit models often used to represent mode choice and discussed in the previous chapter. The systematic trip distribution model is calibrated from the lower hierarchy within segments when between-segment effects are absorbed by the K factors.

A trip distribution model can also be calibrated from the upper level of the hierarchy between segments, but this needs an aggregate measure of cost between segments. The effect of using alternative spatial units such as zones or sectors is an issue in spatial modelling known as the modifiable areal unit problem (MAUP).

The effects of aggregation to progressively larger sectors were also explored in the random model. Two sets of hypothetical spatial error patterns were considered, block and continuous, corresponding with the top-down and bottom-up approaches. The geospatial method of ‘regularisation’ was used to calculate covariance patterns between segments of different sizes.

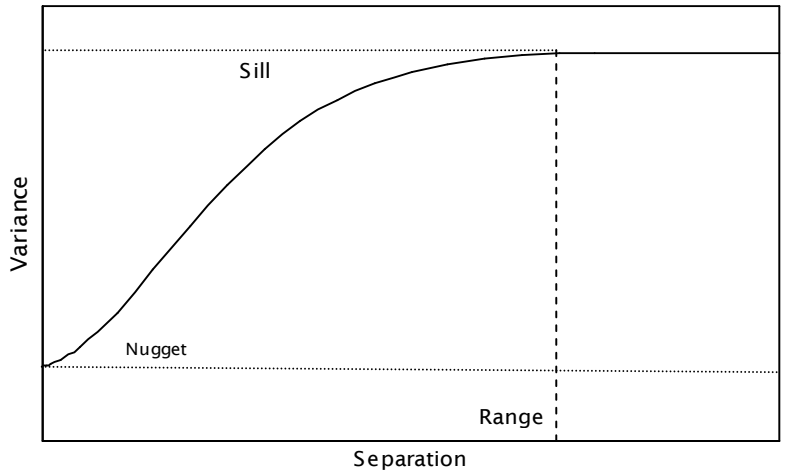
While the conventional approach to geospatial statistics is founded on normal errors, the ‘lorelogram’ works from the log-odds ratio of binary variables. This allows the presence of observations to be processed at the zonal level, without need for aggregation to sectors, to reveal spatial correlations graphically. However, simple processing does not allow for the systematic effects of trip distribution, which has been demonstrated to produce the appearance of correlation as an aspect of Simpson’s paradox.

7.2 Geospatial theory

Much of the early development of geospatial statistics was in the context of mining and geology by Matheron and Krige, and also in forestry by Matérn. It retains some of the terminology of those fields but has since developed into a branch of statistics with many fields of application. The primary references for this research have been Diggle and Ribeiro (2007) and Webster and Oliver (2007). Geospatial analysis is similar to time series analysis, but associates correlation with separation in space rather than time and focuses on interpolation in two or three dimensions rather than extrapolation into the future.

The basic theory is that there is a constant 'sill' variance between widely spaced samples, but the variance is reduced between samples within a certain 'range'; this can be seen as a correlation between them. The relationship between variance and separation can be plotted as a 'variogram', and fitted to various forms of curve.

Figure 7.1 Example of variogram



While no particular form of curve is expected from simple theory, some forms of curve are inconsistent with it. For example, a simple step function with no variance up to 1km and a constant variance beyond that cannot be consistent. If points A, B and C lie along a straight line at ¾km intervals, then samples at A and C must both be equal to one at B, being less than 1km away. However, the samples at A and C should also differ by the variance, being more than 1km apart.

Variogram curves which avoid such inconsistency are known as 'authorised'; the limiting case is a parabola. A simple authorised curve is the exponential, adopted in this study for simplicity and similarity with the deterrence function. It is common to two major sets of authorised curve, the Matérn and the powered exponential families.

Once a form of correlation is determined from a variogram, it can be used to predict values at unsampled points from the values at sampled points. The best estimator can be a linear function of observations within range. The weights sum to unity and tend to decrease with distance, but are affected by clustering of sample points and can even be negative for sample points 'shaded' by other sample points. This calculation of weights and predictions is known as 'kriging'.

Relationships depend on the area, or 'support', over which a sample is taken. The basic theory addresses punctual support at a point, and any residual variance at a single point is known as 'nugget' variance. Without nugget variance, the prediction at a sampled point must be the value of that sample alone,

producing spikes at the sample points in a prediction surface. The extension from punctual to area or zonal support is known as 'regularisation'.

As in most branches of statistics, the simplest and most extensive theory is based on the normal distribution for errors. It also depends on the random process, in particular its underlying mean, not varying in space, 'stationarity'. Fotheringham et al (2002) have developed geographically weighted regression (GWR), in which model coefficients vary spatially.

7.3 Mechanisms of spatial correlation

Geospatial texts recognise that there may be no intrinsic mechanism for spatial correlation. It is likely to appear as an artefact of incomplete model specification, where the omitted or unobserved explanatory variable has its own spatial distribution.

7.3.1 Socio-economic

Work trips have been segmented into white- and blue-collar in several transport models, with mixed results. Differences in cost deterrence functions have been associated with socio-economic factors in disaggregate analyses of Wellington data. If the location of homes and of workplaces varies with socio-economic groupings, these land-use patterns or differences in willingness to travel could appear as a spatial correlation in a single trip distribution.

7.3.2 Modal

This study is based on car trips. Competition from rail services for certain movements might induce a spatial correlation in the distribution of car trips. Some main effects of public transport accessibility are omitted by calibrating against car trip ends, and the topography of the Wellington region has channelled railways and motorways into the same development corridors. Even in a multi-modal distribution, any mis-specification of the joint cost or inclusive value (eg logsum) across modes could induce some similar correlation.

In a strategic model, slow modes (walk and cycle) may have trip lengths on the same scale as zone size. In transport terms, such trips may tend to be intrazonal, and in geospatial terms their covariance may appear as nugget – awkward aspects of the respective models.

7.3.3 Temporal – long term

Growth, movement and decay of different types of industry and employment may leave patterns in the trip distribution that do not match current land uses. Similarly, commuting from newer residential areas may still reflect patterns of when they were developed. Such lag effects might appear in land-use models, but not conventional transport models.

7.3.4 Temporal – short term

Peaking and congestion patterns in different parts of the network, or the schedule of competing rail services, may produce spatial effects when set against desired arrival and departure times.

7.3.5 Network

Trip distribution is a function of costs through the modelled highway network. An error in estimating the cost of one link in the network would induce a correlation between movements using the link, which could appear as a spatial correlation in the trip distribution. Errors could arise in estimating delays due to

congestion, or in their perceived deterrence, like waiting and walking time for public transport. Different elements of the network, as extensive as motorways or as individual as right turns, might be perceived as being greater or lesser deterrents.

Correlation through common links is a consideration in route choice logit models such as Bioroute. This network approach would probably be very demanding in preparation and computation on the scale of the full WTSM model.

7.4 Equivalence of mixed logit and hierarchical forms

When area-to-area K factors are introduced as random terms they are equivalent to the error component formulation of mixed logit that was examined in chapter 6.

The example in that chapter considered three options/alternatives in a simple nesting structure. These could have been modes representing the classic red bus/blue bus problem, or an upper nest between public and private modes.

Table 7.1 Nested mode choice as mixed logit

Mode	A	B	C
Individual	**Train**	**Red bus**	**Blue bus**
h			
i	v_A	v_{BC}	
j			

The random error component is v in HGLM terminology. There is a common value v_{BC} for modes B and C, which represents the common variation or correlation in their utility (subscript for individual i is omitted).

Table 7.2 Hierarchical trip distribution as mixed logit, with random K factors

Sector	Attr	A				B				C				D…
Prod	**Zone**	**11**	**12**	**13**	**14**	**21**	**22**	**23**	**24**	**31**	**32**	**33**	**34**	**~**
H	:													~
I	51	K_{IA}				K_{IB}				K_{IC}				~
	52													
	53													
J	:													~

A similar structure occurs in a sub-rectangle within a PA matrix. Instead of v, K factors can now be seen as the random error component. They represent common variation or correlation between attraction zones 11…14, 21…24, 31…34 in sectors A, B, C…

They also represent commonality or correlation between the zones in sector I at the production end of the movement. Production sectors could be redefined as individual zones, households, persons or trips as has been done in disaggregate modelling.

This interpretation of random-term K factors as a mixed logit model also leads to a hierarchical nesting structure, as in the mode-split example. Hierarchical trip distribution has precedents in the UK National Transport Model and a model of Norwegian by Hamre et al (2002). The calibration of sub-sets within a hierarchy may also be the purpose of Spiess' BalZ3x3 macro for EMME/2. See www.inro.ca/en/download/public/share/macros/balz3x3.mac

7.5 Data preparation – segments and sectors

The top-down approach is to fit K factors to a few large segments and progressively increase the number of factors with smaller segments.

Segments refer to groupings of production–attraction movements – a part of the matrix. In this chapter, all segments are defined by **sectors**, groupings of zones – a part of the study area. All segmentations used here are 'square' in that both productions and attractions are aggregated on the same set of sectors, eg 3x3 or 15x15, but not 3x15.

The WTSM has defined groupings of 15 internal sectors. Aggregations of these have been used to define three, six and 10-sector sets. The three sectors are broadly:

1 Wellington City including the *CBD*

2 SH1: Johnsonville, Porirua and *Kapiti Coast*

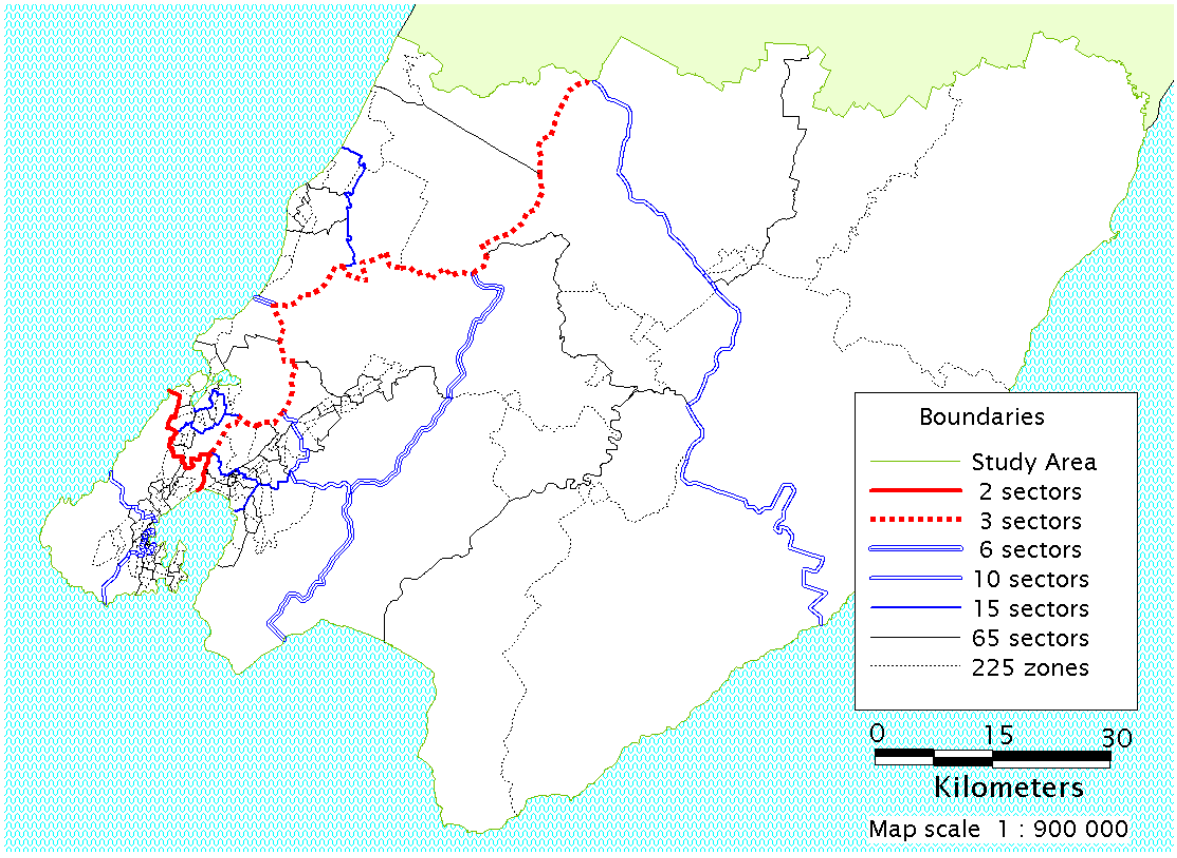
3 SH2: Hutt Valley and *Wairarapa.*

The six sectors separate the areas italicised above. The 10 sectors are nested within the six sectors, so there is a complete nesting hierarchy, ie all boundaries in smaller sets appear in larger sets. Because Tawa is distinctly in the SH1 corridor, only the 15-sector grouping includes all the boundaries of territorial local authorities, which is used as a higher grouping in the WTSM.

A further set of 65 sectors was formed as an intermediate level between the 15 WTSM sectors and the 225 internal zones. It was initially formed by grouping zones with common expansion factors in the HIS. The sectors were then adjusted manually. Groupings with some observed trip ends, either production or attraction, were preferred. All sectors have some observed attraction trip ends, but seven sectors have no production trip ends – five in the Wellington CBD and two in Lower Hutt. There are between two and six zones in a sector. The minimum of two ensures a sector will not be aliased with a zone. Some isolated attraction zones, such as Seaview, might have been suitable sectors otherwise. There are between two and eight such sectors in the 15 WTSM sectors.

A hierarchical numbering system, such as that shown in figure 7.7, was adopted. Numbering is generally outwards from the centre of Wellington.

Figure 7.2 Sector system



7.5.1 Sparse and empty segments

As segments become smaller, the number and proportion of segments with few or no observed trips increase rapidly.

Table 7.3 Sparse and empty segments

Segmentation (production x attraction)	Number of segments				
	Total	Sparse < 10 observed trips		Empty no observed trips	
3 x 3	9	0	0%	0	0%
6 x 6	36	17	47%	7	19%
10 x 10	100	58	58%	30	30%
15 x 15 (WTSM)	225	169	75%	94	42%
65 x 65	4225	4166	98.6%	3490	83%
225 x 225 (zones)	50,625	50,604	99.96%	49,382	97%
– due to empty zones				*19,197*	*38%*

HBW weekday trips by car, from the WTSM HIS (2001)

Empty segments with no observed trips appear even at the coarse 6x6 segmentation, mainly due to the small number of productions from the Wellington CBD.

K factors fitted as fixed effects to empty segments will be zero (or very small). This may exaggerate the fit achieved by the K factors. Empty segments can be omitted from analysis in the same way as empty zones (section 3.8) but this complicates the comparison of different segmentations, particularly in the presence of empty zones with no trip ends. There are also seven empty production sectors, in the Wellington and Lower Hutt CBDs, in the 65-sector system.

K factors treated as random effects did not show any obvious ill-conditioning from fitting to empty segments, possibly because they are also expected to fit the random error distribution.

7.5.2 Intrasector trips

As the size of sector increases, the proportion of intrasector trips increases.

Table 7.4 Intrasector trips

Number of sectors	225	65	15	10	6	3
Observed count	9%	18%	41%	52%	62%	76%
Expanded trips	9%	18%	40%	52%	62%	78%
Trips in Exponential model	6%	15%	36%	48%	60%	78%
Trips in Tanner model	9%	18%	39%	50%	60%	77%

Fitted trip distribution models match observations well in these measures, particularly with the Tanner deterrence function.

7.6 Aggregation and the modifiable areal unit problem

Fitting K factors for each segment absorbs information about contrasts between them; the deterrence function must then be calibrated largely from the within-segment contrasts. This suggests a complementary calibration from between-segment contrasts, simply treating aggregated sectors as larger zones.

While trips can simply be summed when aggregating to larger units, some form of averaging has to be applied to the costs between them. Ideally, the method would be such that the calibration of any aggregation of a synthesised trip distribution returns the same parameters from which it was generated.

The effect of different scales and patterns of zoning has been termed the modifiable (or multiple) areal unit problem (MAUP) by Openshaw (1984). It is an area of continuing research, by the UK Office for National Statistics among others. Fotheringham and Wong (1991) found that the results of multivariate analysis could vary dramatically with different zoning systems. Putman and Chung (1989) modelled housing from several variables such as vacant land and found that rational groupings gave better results. They related housing to employment by a gravity model, but did not describe the definition or aggregation of cost for different zoning systems.

7.6.1 Theoretical

On consideration, it appears that there can be no general method of aggregating costs that gives consistent calibration of a trip distribution, because:

- diverse balancing factors cannot be accommodated

- while the usual grouping would be of contiguous zones, all with similar separations from other zones, a general method would have to apply to any 'grouping', possibly of zones selected from the four corners of the study area.

There may be trivial methods of aggregation that rely on the parameters that are to be recovered in calibration.

There may be special cases where consistent aggregation is possible. An obvious one is the aggregation of a set of zones whose costs to all other zones are identical. This has already been explored in the disaggregation of data units from zones to households, persons and trips, where the same zone-to-zone costs were applied throughout.

A more complex case may be linearity in costs, as where all zones lie along one road or represent different floors served by a lift. The cost-minimising solution to the 'transport problem' of operational research is indeterminate, because shadow costs are equal. Shadow costs play a similar role to balancing factors in trip distribution.

Cases close to linearity may allow aggregations of cost as a good approximation. This line of reasoning supports the selection of the SH1 and SH2 corridors as sectors.

Closeness to linearity might be explored by multi-dimensional decomposition methods, such as principal component analysis, looking for dominant first components. A quick run of the whole of the Wellington cost matrix through multi-dimensional scaling suggests that it can be described by five principal components, compared with the two dimensions needed for simple crow-fly distance. With indirect routes and a time component affected by link speed and congestion, the cost network will not 'lie flat' in two dimensions – another three are needed for a good fit.

7.6.2 Empirical

Empirical tests were made on a variety of aggregation methods. The test case was an aggregation of the Wellington HBW car trip matrix from its original 225 internal zones to just three sectors:

1 Wellington City including CBD

2 SH1: Johnsonville, Porirua and Kapiti Coast

3 SH2: Hutt Valley and Wairarapa.

This gives nine segments, with 79% of observed trips in the three intrasector segments. With five balancing factors (including an overall constant term), there are just four degrees of freedom in which to fit the effects of cost on trip distribution.

Two methods of aggregating costs were considered; the arithmetic mean and the logsum derived from choice theory. The calculation of the logsum requires the cost coefficient from the deterrence function (0.06378 in this instance), implying knowledge of the underlying model.

Five weighting schemes were considered. The simplest was weighting equally by matrix cell, making no allowance for different zone sizes. The second allowed for this with weighting by the observed zonal trip ends, both production and attraction, in effect a 'flat' proportional distribution model taking no account of costs. Costs were incorporated in the synthesised trips from an Exponential model taken as the third weighting scheme. The fourth weighting was by trips synthesised by the generally better-fitting Tanner (gamma) model. Both these synthesised trip weightings incorporate information from a calibration. This is avoided in the final weighting by observed trips. However, for finer aggregations there may be no observed trips in some segments, so their cost by this weighting scheme is undefined.

Table 7.5 Cost aggregation schemes

Weighting	Cost (generalised minutes)			Coefficient fitted to	
	Intra-city	SH2>SH1	Global	Synthesised	Observed
Simple mean					
Cell	18.55	92.72	65.88	0.0538	0.0558
Flat distribution	19.56	100.56	67.76	0.0521	0.0541
Fitted trips (Exponential)	16.07	52.71	22.70	0.0763	0.0800
Fitted trips (Tanner)	15.02	57.29	22.70	0.0667	0.0698
Observed trips	14.89	63.60	22.70	0.0558	0.0582
Logsum					
Cell	15.51	60.09	33.31	0.0703	0.0731
Flat distribution	17.21	63.53	33.17	0.0716	0.0746
Fitted trips (Exponential)	14.13	49.49	16.41	0.0804	0.0847
Fitted trips (Tanner)	12.97	51.61	15.37	0.0728	0.0765
Observed trips	12.93	57.74	15.26	0.0687	0.0724

Table 7.5 compares results from these aggregation methods. The first two columns of figures are aggregated costs for the segments with the largest and smallest numbers of observed trips, the largest being intrasector movements within the city sector. The third cost is the global average and should be the same for any intermediate aggregation by the same method.

Costs weighted by cells and by flat distribution are similar, suggesting a reasonable spread of trip ends between zones. These costs are consistently higher than those weighted by trips because trip distribution favours shorter trips. Logsum costs are lower because they favour lower costs and include a benefit of choice. The simple mean of the global cost weighted by observed trips is the mean trip cost, 22.70 generalised minutes, which is replicated in the model fittings.

The table also shows the coefficients of cost fitted in simple Exponential models of synthesised and observed trips aggregated by summation to the 3×3 level. The synthesised trips are from a model fitted at the zonal level with a coefficient of 0.06378. Under a consistent aggregation scheme this would be replicated in the aggregate model, but none of the tested schemes do so. Their coefficients are still reasonably close to this value and might just be acceptable for practical purposes.

Cost coefficients for observed trips are all larger by broadly similar amounts (3.6%–5.4%). This suggests a hierarchical effect, which might be:

- K factors as random error components leading to a nested model
- Fotheringham's spatial information processing
- an artefact of mis-specifying a simple Exponential deterrence function where the Tanner fits better.

Significances of fitting the aggregated costs are high, but are not readily comparable.

The cost coefficient fitted to observed trips aggregated to 3×3 segments by the model estimation methods of the next chapter ranges from 0.068 to 0.073, depending on the weight given to trip ends (table 8.17, top line).

None of the obvious methods of cost aggregation tried here give the desired consistency. However, all provide reasonable model fitting despite gross aggregation. Calibration of synthesised trips might provide a useful baseline to account for aggregation effects.

7.6.3 Practical

In practice, models built at different scales will have different networks from which costs are derived. Coarser models will only include major roads and their zones are likely to be loaded at major intersections, perhaps the central crossroads of a community. In finer models, zones may be loaded onto links between junctions, or side roads, so appropriate turning movements are assigned to the junctions for their modelling, as in SATURN.

As zone sizes grow, the length of the centroid connectors may increase. The costs on centroid connectors may be added to zone-to-zone costs, but they represent the variability in costs to and from different parts of the zone more than a constant addition to them. As such, they might be included in a dispersion model of an HGLM, or a mixing term in a mixed logit model.

7.6.4 Solution

A solution to calibrating a trip distribution from between-segment information was found in calibration to aggregate data, using matrix estimation methods as described in chapter 8. This did not address the MAUP of aggregating costs, but circumvented it by modelling a full zonal matrix with zone-to-zone costs and fitting it to totals of observed trips for whole segments.

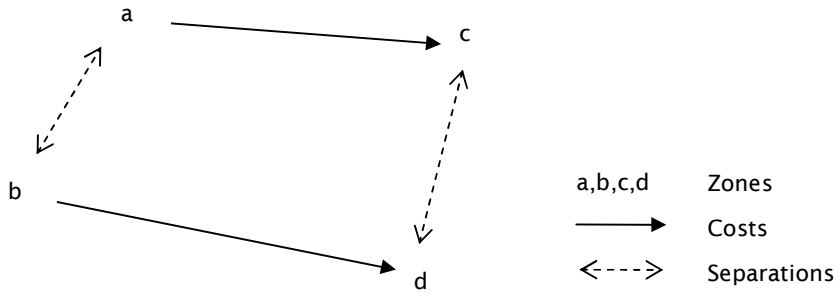
7.7 Separations and costs

The bottom-up approach relates correlations in the errors from fitting a trip distribution model to separation. Separation can be expressed simply as distance. In some studies an ecological closeness is measured by the number of species in common; there may be similar measures of socio-economic separation. Measures of separation can also be formed from the adjacency of zones, but many common boundaries between WTSM zones lie along mountain ranges that actually separate the communities to either side and do not suggest similarity between them.

Transport models offer a ready-made measure of separation in costs measured along the transport network. These recognise the topography of the study area, incorporating the separation imposed by sea inlets and mountain ranges, which effectively divide zones that are physically close or even adjacent. On reflection, the closeness in generalised cost provided by motorways or other high-speed links may not be matched by a similarity in choice of workplace; distances along the network may be better proxies for social similarities.

Correlations are sought in the numbers of trips for similar production–attraction movements. Unlike most subjects of geospatial analysis which are points or areas, production–attraction movements are defined by two areas (zones), which may be simplified to two points (zone centroids). This leads to two measures of separation; the separation of the production ends and the separation of the attraction ends. For this study, these two separations have simply been averaged. Other combinations such as a Pythagorean root mean square are possible and there are methods for analysis of separations in multiple dimensions, more usually spatial dimensions such as eastings and northings in anisotropic models.

Figure 7.3 Map of costs and separations



Generalised costs are asymmetrical; the cost from P to A is not identical to the cost from A to P. This arises from one-way systems, turning penalties in modelled junctions, congestion in the peak direction and parking charges in the CBD. There is no logical reason for asymmetry in the separations between two production–attraction movements, so separations were made symmetrical by removing parking charges and simply averaging the costs over the two directions.

Intrazonal costs were recalculated after CBD parking charges were omitted. The five-minute maximum was not imposed to allow the intrazonal cost to reflect more the size of zone, giving a lower correlation between movements to or from larger zones.

7.7.1 Duality of separation and cost

One reason for choosing network costs as a measure of separation was a duality between them.

Setting asymmetrical effects aside, trips for movements a↔c and b↔d are deterred by costs $C_{a\leftrightarrow c}$ and $C_{b\leftrightarrow d}$, and separated by $C_{a\leftrightarrow b}$ and $C_{c\leftrightarrow d}$ which may determine a degree of correlation. For movements a↔b and c↔d, the roles of costs in deterrence and correlation are exchanged. Both deterrence and correlation can be modelled as Exponential functions of network cost.

Figure 7.4 Matrix of costs and separations

Zones	~	**a**	~	**b**	~	~	**c**	~	**d**	~
:										
a				S_p			C		X	
:										
b		S_p					X		C	
:										
:										
c		c		x					S_a	
:										
d		x		c			S_a			
:										

Key

C = costs determining deterrence in trip distribution

S = separations determining spatial correlation

$_{p,a}$ = subscripts for production and attraction separations

Pairs of movements whose correlations are determined by the same separations:

X = crossed

c = reversed

x = reversed and crossed

assuming symmetry and isotropy.

Note that the same separations apply to the 'crossed' movements a↔d and b↔c, assuming symmetry in attraction and production separations, and to the reverse movements c↔a and d↔b, assuming isotropy between attraction and production separations. The latter can be seen as a transposition of the matrix.

Strictly speaking, network costs refer to origin-destination movements, and are specific to time periods. Production–attraction costs are averages of these, weighted by the typical time and direction of travel, ie with peak flows in peak periods for HBW. This distinction becomes important when modelling time period choice.

It was hoped that some further insight into the process or reduction of the problem would develop from this duality, but none was recognised in the course of the study. The term separation is used to distinguish the determinant of correlation effects, with its slightly different symmetric formulation, from the costs in the deterrence function of trip distribution.

7.8 Regularisation

Computing limitations restrict both top-down and bottom-up analyses to quite high levels of aggregation. To investigate the effect of the limitations, idealised spatial error patterns were aggregated to different levels. The aggregation process is known as regularisation. It is not a simulation process like the randomisations used to investigate mixed models, but is based on the analytical properties of variances and covariances. The main results are shown in figures 7.8, 7.9 and 7.10, and tables 7.6 and 7.7.

Two sets of spatial error patterns were considered, block and continuous, broadly corresponding to the top-down and bottom-up approaches. In the block patterns, there is a unit variance for each segment and no correlation between segments at a given level of aggregation to sectors. This level of aggregation is termed the block error level; it is distinguished by plotting characters in figure 7.8 and by different columns in table 7.6. A set of patterns is generated from different blocking levels. This corresponds with the top-down approach of fitting K factors as random terms.

In continuous error patterns, the correlations are a function of separation, specifically an exponential function:

$$\rho = \exp(-\phi \times \text{separation})$$

as fitted in the bottom-up approach. A set of patterns is generated by different values of ϕ. The inverse of ϕ is proportional to the range of correlation effects, with the dimension of generalised minutes. In this study, it is termed ‘characteristic separation’.

The exponential function does not have a finite range, but the practical range is three times the characteristic separation (where $\rho = 0.05$, after Diggle and Ribbeiro (2007, section 3.4.1, p51); $0.05 \approx \exp(-3)$).

7.8.1 Calculation

The form of the trip distribution model is multiplicative, including the K factors which are treated here as random error terms. Hence trips T are distributed:

$$T \sim T_{mean} \times \varepsilon$$

where ε has a mean of 1 and a variance of σ^2

thus $$Var(T) = T^2\sigma^2$$

and $$Covar(T_i,T_j) = T_iT_j\sigma_i\sigma_j\rho_{ij}$$

Consider an aggregation from smaller segments i,j to larger segments I,J. Note that these indices represent segments, ie zone-to-zone or sector-to-sector movements, and not zones or sectors. Trips in the larger segments are simply summed from the smaller segments they include:

$$T_I = \Sigma_{i \in I} T_i$$

and from basic properties of variance and covariances:

$$\text{Var}(T_I) = \Sigma_{i\in I}\Sigma_{j\in I}\,\text{Cov}(T_i,T_j) \qquad \text{including } \text{Cov}(T_i,T_i) = \text{Var}(T_i)$$

$$= \Sigma_{i\in I}\Sigma_{j\in I}\,T_iT_j\sigma^2\rho_{ij}$$

$$= \sigma^2\Sigma_{i\in I}\Sigma_{j\in I}\,T_iT_j\rho_{ij} \qquad \text{ie weighted by trips.}$$

Similarly $$\text{Covar}(T_I,T_J) = \Sigma_{i\in I}\Sigma_{j\in J}\,\text{Cov}(T_i,T_j)$$

$$= \sigma^2\Sigma_{i\in I}\Sigma_{j\in J}\,T_iT_j\rho_{ij}$$

By analogy, a correlation ρ_{IJ} between the larger segments can be defined as the average of the movement correlations, weighted by the product or square of trips:

$$T_IT_J\sigma^2\rho_{IJ} = \Sigma_{i\in I}\Sigma_{j\in J}\,T_iT_j\sigma^2\rho_{ij}$$

$$\rho_{IJ} = \Sigma_{i\in I}\Sigma_{j\in J}\,T_iT_j\rho_{ij} \;/\; \Sigma_{i\in I}T_i\,\Sigma_{j\in J}\,T_j$$

$$= \Sigma_{i\in I}\Sigma_{j\in J}\,T_iT_j\rho_{ij} \;/\; \Sigma_{i\in I}\Sigma_{j\in J}\,T_iT_j$$

The relationship $\rho_{IJ}=f(\text{Separation}_{IJ})$ is no longer exact but by using the same weighting to define segment separations

$$\text{Separation}_{IJ} = \Sigma_{i\in I}\Sigma_{j\in J}\,T_iT_j\text{Separation}_{ij} \;/\; \Sigma_{i\in I}\Sigma_{j\in J}\,T_iT_j$$

the relationship is generally a good approximation when the linearisation is over a limited range of separations between segments I and J. The smaller the range, the less the weighting scheme matters and the better the approximation to the original correlation function.

However, this analogous measure ρ_{IJ} for larger segments is not a correlation in the conventional sense, since self-correlation is not necessarily equal to unity

$$\rho_{II} \neq 1$$

because it is a weighted average of ρ_{ij} which is less than or equal to unity. It can be normalised by redefining the variance and correlation of the larger segments as σ'^2 and ρ'

$$\sigma'^2_I = \sigma^2\rho_{II}$$

$$\rho'_{IJ} = \rho_{IJ}\,\sigma^2 / \sigma'_I\,\sigma'_J$$

This rescaling of variance with block size is the basis for Krige's relationship, which decomposes the variance in a region into between-block and within-block variances, and also for the nested sampling and analysis used in the top-down approach.

Regularisation typically relates variations between square blocks of land or ore to those observed between point samples. Equal-sized blocks and a stationary punctual error process lead to a common variance for all blocks, $\sigma'^2=\sigma'^2_I$ for all I.

In this study, the sectors are irregularly sized and shaped, so there is no common variance for the segments based on them. Moreover, there are several different levels of such irregular aggregation to be compared. For ease of comparison, all error patterns are related to a common unit variance, setting σ=1.

For blocked patterns, this unit variance applies to the segments at the level of the blocking. For continuous patterns, a unit punctual variance is assumed; this is related to errors at the zonal level by including appropriate intrazonal costs in the separation when calculating correlations. Thus for a minimal segment comprising the internal movements within one zone

$$\text{variance of intrazonal trips} = \exp(-\phi \times \text{intrazonal cost})$$

Other formulations for intrazonal separation could be considered, as in the comparison of intrazonal cost formulations in section 4.11, but this provides a simple, basic allowance for varying zone sizes. When used as a separation, intrazonal costs are not capped at five generalised minutes, allowing greater internal effects within larger zones.

The error patterns at the different aggregation levels are expressed relative to these block or punctual unit variances, as variances and covariances.

These variances and covariances are scaled by ratios of trips. The terms T_I^2 for variances or $T_I T_J$ for covariances in the equations above thus act as weights in averaging these measures of error for aggregation. For this analysis, trips fitted to a trip distribution model with an Exponential deterrence function have been used for this weighting.

7.8.2 Aggregation into blocks

The error patterns and their aggregations can be seen in the structure of a correlation or variance-covariance matrix. The correlations are between movements, each defined by a pair of zones or sectors. Each cell of a trip or cost matrix appears as a row or column of a correlation matrix. The matrix is thus much larger than a trip or cost matrix.

Figure 7.5 sets out a correlation matrix for four zones. The rows and columns are ordered by the production and attraction zones. These zones are aggregated to two sectors, comprising zones 1 and 2, and zones 3 and 4, giving four segments:

Segment	Production zones	Attraction zones
A	1 & 2	1 & 2
B	1 & 2	3 & 4
C	3 & 4	1 & 2
D	3 & 4	3 & 4

Thus $T_A = T_{11} + T_{12} + T_{21} + T_{22}$

Since $\text{Var}(T_A) = \Sigma_{i \in A} \Sigma_{j \in A} \text{Cov}(T_i, T_j)$ including $\text{Cov}(T_i, T_i) = \text{Var}(T_i)$

and indexing the right-hand side of the equations by the production-attraction zone numbers

then $\text{Var}(T_A) = \text{Var}(T_{11}) + \text{Cov}(T_{11},T_{12}) + \text{Cov}(T_{11},T_{21}) + \text{Cov}(T_{11},T_{22})$

$+ \text{Cov}(T_{12},T_{11}) + \text{Var}(T_{12}) + \text{Cov}(T_{12},T_{21}) + \text{Cov}(T_{12},T_{22})$

$+ \text{Cov}(T_{21},T_{11}) + \text{Cov}(T_{21},T_{12}) + \text{Var}(T_{21}) + \text{Cov}(T_{21},T_{22})$

$+ \text{Cov}(T_{22},T_{11}) + \text{Cov}(T_{22},T_{12}) + \text{Cov}(T_{22},T_{21}) + \text{Var}(T_{22})$

Noting that $\text{Cov}(X,Y) = \text{Cov}(Y,X)$, this reduces to the familiar form

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$$

Thus segment variances comprise variances of zone-to-zone movements, and some of the covariances. In figure 7.5 the black cells on the diagonal hold the variances for zone-to-zone movements. The covariances between zone-to-zone movements that contribute to the segment variances are shown in grey.

Figure 7.5 Covariance matrix ordered by zone-to-zone movements

Segment			A	A	B	B	A	A	B	B	C	C	D	D	C	C	D	D
	Zone	Prod	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
	Prod	Attr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
A	1	1																
A	1	2																
B	1	3																
B	1	4																
A	2	1																
A	2	2																
B	2	3																
B	2	4																
C	3	1																
C	3	2																
D	3	3																
D	3	4																
C	4	1																
C	4	2																
D	4	3																
D	4	4																

Black = zonal movement variances; grey = contributions to segment level variances

Reordering the matrix rows and columns by segment reveals a block structure in figure 7.6.

Figure 7.6 Covariance matrix ordered by segments

Segment			A	A	A	A	B	B	B	B	C	C	C	C	D	D	D	D
	Zone	Prod	1	1	2	2	1	1	2	2	3	3	4	4	3	3	4	4
	Prod	Attr	1	2	1	2	3	4	3	4	1	2	1	2	3	4	3	4
A	1	1																
A	1	2																
A	2	1																
A	2	2																
B	1	3																
B	1	4																
B	2	3																
B	2	4																
C	3	1																
C	3	2																
C	4	1																
C	4	2																
D	3	3																
D	3	4																
D	4	3																
D	4	4																

Figure 7.7 introduces a hierarchical segmentation system, omitting zone pairings. The segment at aggregation level N is indicated by the first N digits in the numbering system.

Level of aggregation N	Index from segment number	Contributors to variance
1 - most aggregate	n~~~	
2	nn~~	
3	nnn~	
4 - least aggregate	nnnn	

Blocks of lighter shades include darker shades.

Figure 7.7 Covariance matrix with hierarchical segments

Segment	1111	1112	1121	1122	1211	1212	1221	1222	2111	2112	2121	2122	2211	2212	2221	2222
1111																
1112																
1121																
1122																
1211																
1212																
1221																
1222																
2111																
2112																
2121																
2122																
2211																
2212																
2221																
2222																

In this example, there are only two sub-segments for each level of hierarchy. Usually there would be at least four, when sectors are each made up of at least two sub-sectors, as above. With different sizes of sectors, the sizes of the blocks would differ, but the structure of nested blocks along the diagonal would remain.

Aggregated segment variances are the weighted average of cells in the blocks on the diagonal with the shade of grey for the aggregation level, including darker shades. Covariances are the weighted averages of similar sized blocks off the diagonal, including lighter shaded cells.

With a common scheme of weighting by trips, T_iT_j, the aggregation can be from either the basic disaggregate cells (eg zone-to-zone movements), or from intermediate aggregations of blocks (eg sector-to-sector movements).

This common weighting scheme can be used to give an overall average variance from all the blocks on the diagonal. The complementary set of blocks off the diagonal gives an overall average covariance. These two

measures are tabulated in the upper and lower bodies of tables 7.6 and 7.7. The ultimate aggregation is to a single segment, represented by a single block on the diagonal covering the whole covariance matrix, giving a single variance over the whole study area due to geospatial effects. This appears in the top line of tables 7.6 and 7.7; the bottom lines are derived from the black individual cells on the diagonal for variances and the complementary off-diagonal cells for covariance.

In geospatial statistics, finite study areas lead to distinctions between regional and theoretical variograms and issues of ergodicity.

7.8.2.1 Aggregate separations

Averages of separation within and between segments have been calculated with the same weighting scheme, from the on- and off-diagonal blocks respectively. These provide the horizontal axis in figures 7.8, 7.9 and 7.10; overall averages are shown on the left of tables 7.6 and 7.7. These average separations are broadly similar to those from weighting by a simple count of zone-to-zone movements (typically ±3 generalised minutes). There are not the wider differences between global costs due to different weighting schemes seen in table 7.5.

7.8.3 Block error pattern

The block structure also applies to the blocked error patterns. At the level of segmentation at which the pattern occurs, all variances and covariances within the diagonal blocks are 1; all other covariances are 0. This also makes all correlations within a block equal to 1, so the same error applies to all movements within a block, eg all Wellington city to Hutt Valley movements +5%, all Hutt Valley to Kapiti Coast movements -3%.

If this process is analysed at a less aggregate level than that at which the block error pattern occurs, aggregation will be to smaller blocks. Variances on the diagonal will all be 1; covariances off the diagonal will be 1 where they still fall within the blocks of the error pattern, or 0 outside.

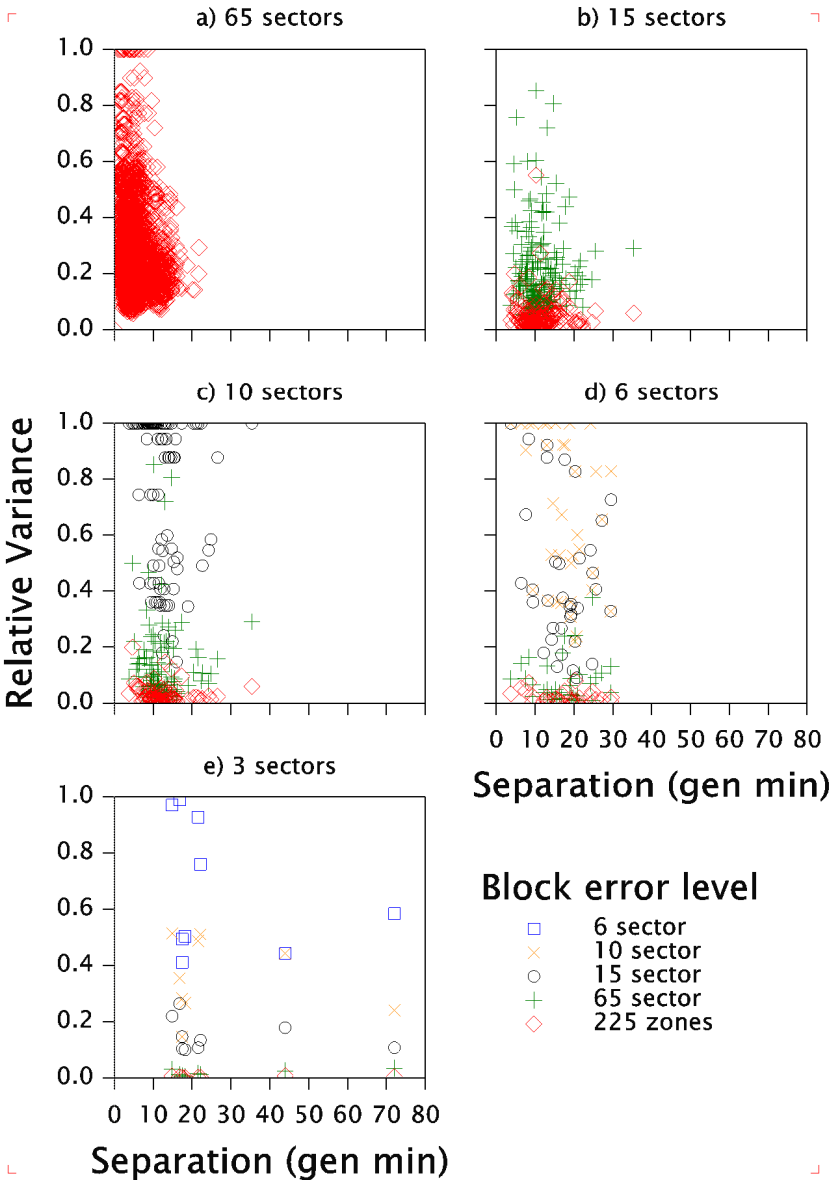
If this process is analysed at a more aggregate level than that at which the block error pattern occurs, aggregation will be to larger blocks. All covariances will fall outside the error pattern blocks, and be 0. Variances will be a weighted average of 1 from the diagonal error pattern blocks, and 0 outside them but still within the diagonal blocks of the analysis aggregation.

Figure 7.8 plots relative variances for individual segments. There is a separate sub-plot for each level to which the segments are aggregated. The different levels at which the block error pattern occurs are indicated by separate plot characters. Where these are the same or more aggregate than the level being plotted, all variances are 1 and would appear on the top edge of the plot, but have been omitted.

Thus in plot (a), aggregated on 65 sectors, only the variances from a block error pattern at the zonal level are shown for 65 x 65 segments.

Plot (b), aggregated on 15 sectors, also shows variances from the block error pattern at the 65 sector level for 15 x 15 segments. Variances from more block error patterns appear or come down from the top edge with increasing aggregation.

Figure 7.8 Individual variances from block error patterns



Some variances appear on the top edge of plots (c) and (d) where segments from block patterns at the 15 and 10 sector levels have not been aggregated with any other segment.

Overall, the variances can be seen to fall with increasing aggregation. Without the weighting by trips, this would just depend on the number of segments from the block error level that had been aggregated. Such simple averaging of independent random numbers would appear as bands at $1/\sqrt{n}$ across the plot, where n is the number of segments from the block error level that have been aggregated. Weighting by trips produces a continuous vertical spread.

Although this spread is considerable for each level of block error, the overlap with other levels is not too large to distinguish a fairly distinct band for each level.

Separations within segments tend to increase along the horizontal axis with aggregation.

Average separations within segments and variances, effectively the centroids of the clusters shown in figure 7.8, are shown in the top half of table 7.6. Average separations between segments and covariances are shown in the bottom half.

Table 7.6 Average variances and covariances from block error patterns

	Average	Block error level						
Aggregation	separation	Number of sectors						
number of	within	225 zones	65	15	10	6	3	1
sectors	segments	Variances						
1 – study area	66.56	0.001	0.005	0.03	0.05	0.13	0.24	1.00
3	45.17	0.003	0.021	0.13	0.22	0.54	1.00	1.00
6	20.91	0.005	0.039	0.24	0.41	1.00	1.00	1.00
10	13.03	0.013	0.096	0.59	1.00	1.00	1.00	1.00
15	9.98	0.022	0.162	1.00	1.00	1.00	1.00	1.00
65	7.90	0.137	1.000	1.00	1.00	1.00	1.00	1.00
225 – zones	4.05	1.000	1.000	1.00	1.00	1.00	1.00	1.00
	between segments	Covariances						
3	73.32	0.000	0.000	0.00	0.00	0.00	0.00	1.00
6	73.32	0.000	0.000	0.00	0.00	0.00	0.13	1.00
10	69.52	0.000	0.000	0.00	0.00	0.08	0.20	1.00
15	68.38	0.000	0.000	0.00	0.02	0.10	0.22	1.00
65	66.86	0.000	0.000	0.03	0.05	0.12	0.24	1.00
225 – zones	66.58	0.000	0.005	0.03	0.05	0.13	0.24	1.00

Average separations within segments increase distinctly with the size of sector up to 66.56 generalised minutes for the whole study area taken as a single sector. The increase in average separation between segments in the lower part of the table is much smaller. This may be interpreted as the separations that become internal as sector sizes increase (in the lighter grey cells of figure 7.6), being almost as large as those that remain between segments, on average. This may be an artefact of weighting the average by the product of distributed trips.

Variances for aggregations less than the block error level (bottom-right of the top half) are all unity. Covariances for aggregations more than the block error level (top left of the bottom half) are all zero.

The right-hand column shows the trivial limiting case of the whole study area as a single sector, giving complete correlation between any sub-divisions.

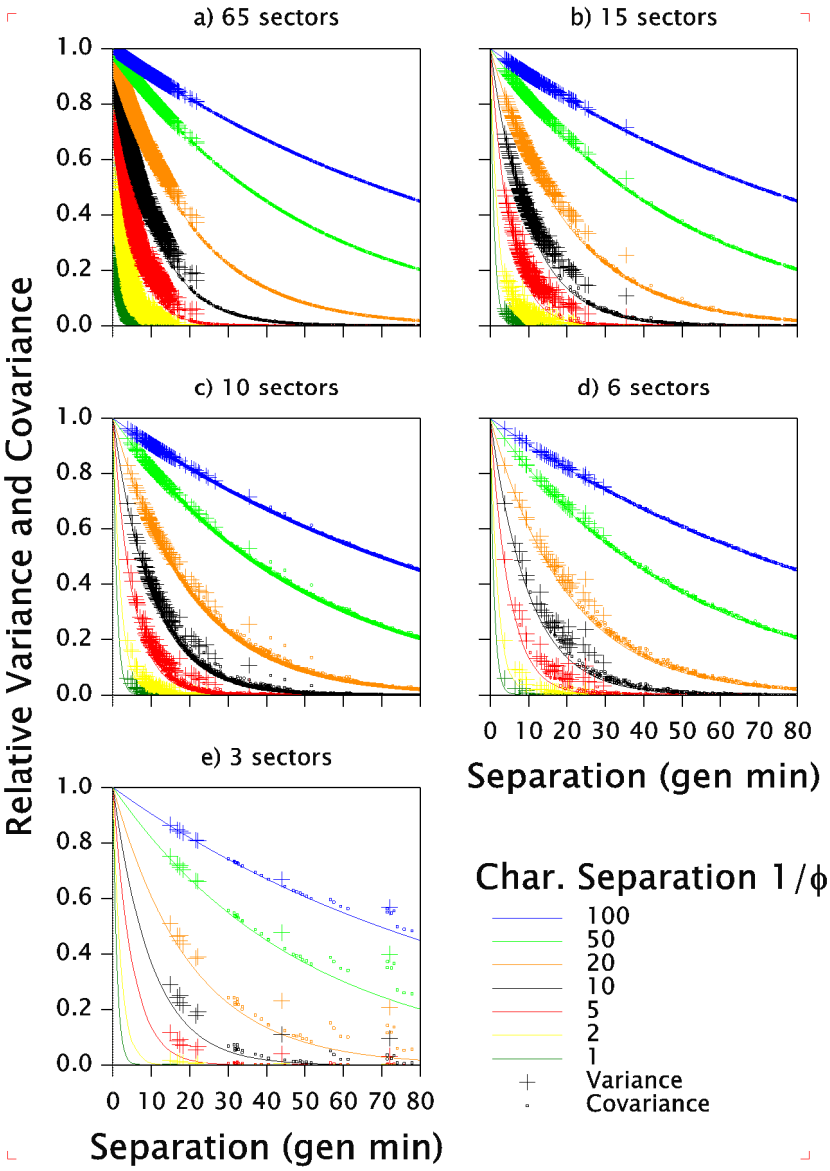
7.8.4 Continuous error patterns

In continuous error patterns, variance is an exponential function of separation between movements. In terms of figure 7.6 correlations are not confined to, and complete within, the grey blocks on the diagonal; there are varying shades of grey across the whole matrix. Even variances of the black diagonal cells representing zone-to-zone movements are less than one because they are the function of internal separation taken from intrazonal costs.

Because individual covariances of all aggregations are either 1 or 0 in block error patterns, they are omitted from figure 7.8. This is not the case for the continuous error patterns plotted in figure 7.9 so

covariances of individual segments are plotted as points. The exponential functions of separation that generate the punctual correlations are plotted as continuous curves.

Figure 7.9 Individual variances and covariances from continuous error patterns



All points lie on or above the generating curves, since they are all linear averages (with varied weighting by trips) of points lying on the curve. At the zonal level of aggregation, all points lie on the curve, but variances are already below the top edge because of incomplete internal correlation represented by $\exp(-\phi \times \text{intrazonal cost})$. This is not plotted because of the very large number of points, but the average effects are shown in the bottom lines of table 7.7. Also to avoid plotting large numbers of points only a sample of covariances, about 0.01% and 3% respectively, are shown in sub-plots (a) and (b).

Table 7.7 Average variances from continuous error patterns

Aggregation number of sectors	Separation within segments	Characteristic separation, $1/\phi$, generalised minutes						
		1	2	5	10	20	50	100
		Variances						
1 – study area	66.56	0.0002	0.002	0.014	0.05	0.16	0.38	0.58
3	45.17	0.0010	0.007	0.052	0.16	0.32	0.55	0.70
6	20.91	0.0016	0.011	0.074	0.21	0.41	0.68	0.82
10	13.03	0.0039	0.025	0.145	0.34	0.55	0.78	0.88
15	9.98	0.0064	0.038	0.197	0.41	0.63	0.82	0.91
65	7.90	0.0228	0.093	0.300	0.51	0.69	0.86	0.93
225 – zones	4.05	0.0473	0.189	0.479	0.68	0.82	0.92	0.96
	between segments	Covariances						
3	73.32	0.0000001	0.00002	0.002	0.02	0.10	0.33	0.54
6	73.32	0.0000213	0.00028	0.005	0.03	0.12	0.34	0.54
10	69.52	0.0000230	0.00034	0.007	0.04	0.13	0.36	0.56
15	68.38	0.0000302	0.00045	0.008	0.04	0.14	0.37	0.57
65	66.86	0.0001141	0.00115	0.013	0.05	0.15	0.38	0.58
225 – zones	66.58	0.0002164	0.00156	0.014	0.05	0.16	0.38	0.58

Average separations are the same as in table 7.6.

The characteristic separation of one generalised minute in the left-hand column is smaller than the typical size of a zone. Correlation effects are mostly averaged out within the zone, leaving a variance of only 0.0473 at the zonal level. This effect diminishes even further with greater aggregation up the column.

In the right-hand column, the characteristic cost of 100 generalised minutes is larger than the average separation over the whole study area. There is relatively little diminution of variances and covariances with aggregation within the study area.

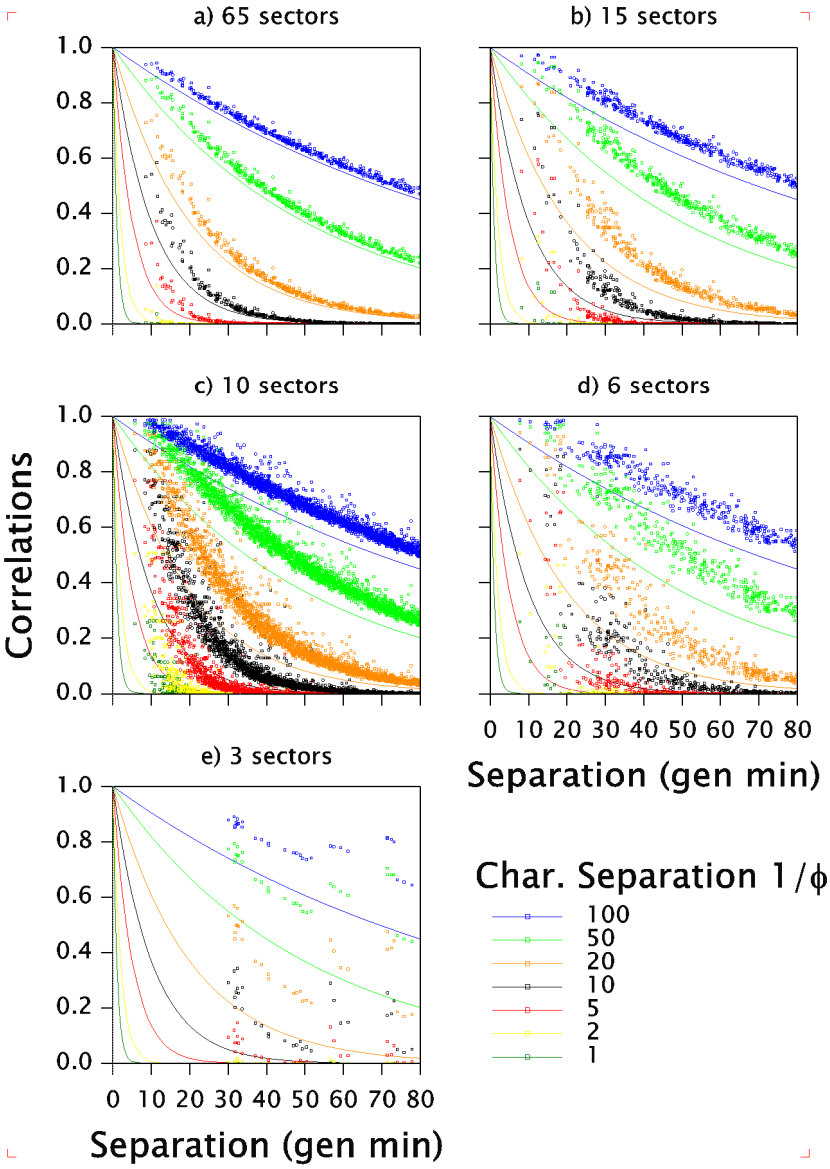
In all cases, covariances are smaller than variances. The proportional differences are smaller for large sectors and large characteristic separations, to the top-right of the table.

The limiting case of increasing characteristic separation, $\phi \to 0$, to the right of table 7.7, is the right-hand column of table 7.6, with unit variances and covariance throughout. The limiting case of decreasing characteristic separation, to the left of table 7.7, is zero variances and covariance throughout. However, if the base unit variance is redefined to apply at the zonal rather than punctual level, this limiting case of $\phi \to \infty$ becomes the left-hand column of table 7.6.

Figure 7.10 shows correlations from continuous error patterns. These are the same datasets as are shown in figure 7.9, but with variances effectively set to unity and covariances normalised accordingly:

$$\text{correlation}_{IJ} = \text{covariance}_{IJ} / \sqrt{(\text{variance}_I \times \text{variance}_J)}$$

Figure 7.10 Correlations from continuous error patterns



Since all relative variances are less than unity, correlations are larger than the relative covariances, and the plots are further above the original generating line for punctual correlation. Despite the increased scatter, the association of correlation with separation is still recognisable.

7.8.5 Discussion

The figures and tables show clearly the marked reductions in relative variances and covariances in spatially correlated error patterns that are apparent with increasing spatial aggregation. Generally, the information available about an effect is proportional to the variance it presents, so the ability to detect short-range effects is greatly diminished in aggregated data. For example, aggregation to the 15 sector level loses all but 16.2% of the information about a block error at 65 segments (table 7.6), or 19.7% from a continuous error with a characteristic separation of five generalised minutes (ϕ=1/5; table 7.7).

The figures also show the marked variability in variance, or heteroscedasticity, due to the irregular size and shape of sectors and distribution of trips between them. This is not accounted for in either the top-down or the original bottom-up approaches. It is likely to obscure the pattern of spatial errors even if a distinct one does exist, though figure 7.10 shows that a relationship between covariance and separation is still apparent after normalisation to homoscedasticity. Allowing general heteroscedasticity in a random error model would be profligate in fitted terms (one per segment) and would abandon any information about the spatial error pattern contained in the heteroscedasticity. Patterns of heteroscedasticity expected from a particular spatial error pattern might be introduced into the random error model as an offset, in a similar way to the introduction of the correlation matrix in the bottom-up approach. As with the correlation matrix, this would be a trial-and-error approach, requiring a good understanding of the measures of fit. Estimating parameters of a spatial error model directly from heteroscedasticty information appears to be very complex.

The process of regularisation highlights the improbable nature of the block error pattern. Its all-or-nothing correlations between segments are inconsistent with any spatial process not defined by the zone or sector boundaries and cannot correspond with any authorised continuous spatial process. Although zone boundaries are chosen to maximise differences between zones and minimise differences within them, they are likely to capture only part of a spatial error process.

A continuous spatial error process can appear as:

- variance between segments – the top-down approach
- heteroscedasticity associated with irregularity
- an association of correlation with separation – the bottom-up approach.

These are different artefacts of the same process and may all contain information about it. Knowledge of the extent to which various patterns of spatial correlation can be detected and distinguished from these artefacts, singly or in combination, would provide a better footing for future studies.

The plots in this exercise are greatly simplified by adopting a common base variance for the block level or punctual error. In fitting models to data, there will be considerable correlation between the estimates for this base variance and for the range of covariance effects as there is between the intercept and slope of a simple regression.

7.9 Top-down approach with K factors

The top-down approach was to fit K factors by area-to-area movements or segments.

K factors are adjustments to the constant in the deterrence function for sets of trips.

They can be treated either as fixed terms in an ordinary GLM, or random terms in a mixed GLM (GLMM or HGLM). Fitting them as fixed terms ensures that synthesised trips match observed trips in each set with a separate K factor, which can be seen as area(-to-area) specific constants in discrete choice modelling terms. As a random term, they affect the accuracy of predictions from the model and give insight into its limitations.

K factors were fitted as random terms to an HGLM of the form

$$f(\mathsf{y}) = \mathsf{X}\beta + \mathsf{Z}\upsilon$$

The systematic part of this model, $f(\mathsf{y}) = \mathsf{X}\beta$, is the same as for previous calibrations of trip distributions by GLM. The data matrix X comprises factors for production and attraction zones, and the travel costs (or natural logarithms of them for Power or Tanner deterrence functions). The fitted coefficients β are the trip end

balancing factors and the cost coefficients for the deterrence function. The link function $f()$ is the logarithm and observations of y are taken to be Poisson distributed, the standard form for a log-linear model.

The additional random component $Z\upsilon$ of the HGLM comprises a factor Z and a random term υ which takes a separate value drawn from a random distribution for each level of Z. The random term υ is taken to be distributed as the logarithm of a gamma distribution, since this is conjugate with the log-Poisson systematic model and thus presents less computational difficulty in model fitting and calculation of statistics.

In this approach, Z is the set of production–attraction segments to which K factors are fitted. HGLM allows several segmentations to be included in a single model, fitting a separate random term υ with its own variance for each segmentation. If K factors were treated as fixed terms, the factors would be aliased because the segmentations are nested.

However, computational limits did not allow the fitting of segmentations finer than 15 x 15, either singly or in combination. Table 7.8 shows the fitted variances of K factors as random terms for all segmentations from 3 x 3 to15 x 15 simultaneously, with either Exponential or Tanner deterrence functions. Convergence of the models was still slow and uncertain.

Table 7.8 Fitted variances for hierarchical K factors

Segmentation	Exponential deterrence function			Tanner deterrence function		
	Variance	t ratio *	Cumulative variance	Variance	t ratio *	Cumulative variance
3x3	0.00006	-0.14	2.11	0.00007	-0.24	1.11
6x6	1.25	0.75	2.11	0.43	-2.29	1.11
10x10	0.25	-4.22	0.86	0.21	-4.95	0.68
15x15	0.61	-3.69	0.61	0.47	-5.39	0.47
65x65	Not included in hierarchy					
Zonal	insufficient computer memory					

* Variances are estimated on the logarithmic scale, and the t ratios are calculated on that scale

This is interpreted as nested sampling and analysis, as set out by Webster and Oliver (2007, section 5.3). HGLM fitting accounts for unequal sampling (Webster and Oliver 2007, section 5.3.3 and table 5.6), allowing stratum variances to be accumulated simply (Webster and Oliver 2007, equation 5.35) to form experimental variograms, shown in figure 7.11.

Figure 7.11 Variogram from accumulated hierarchical K-factors

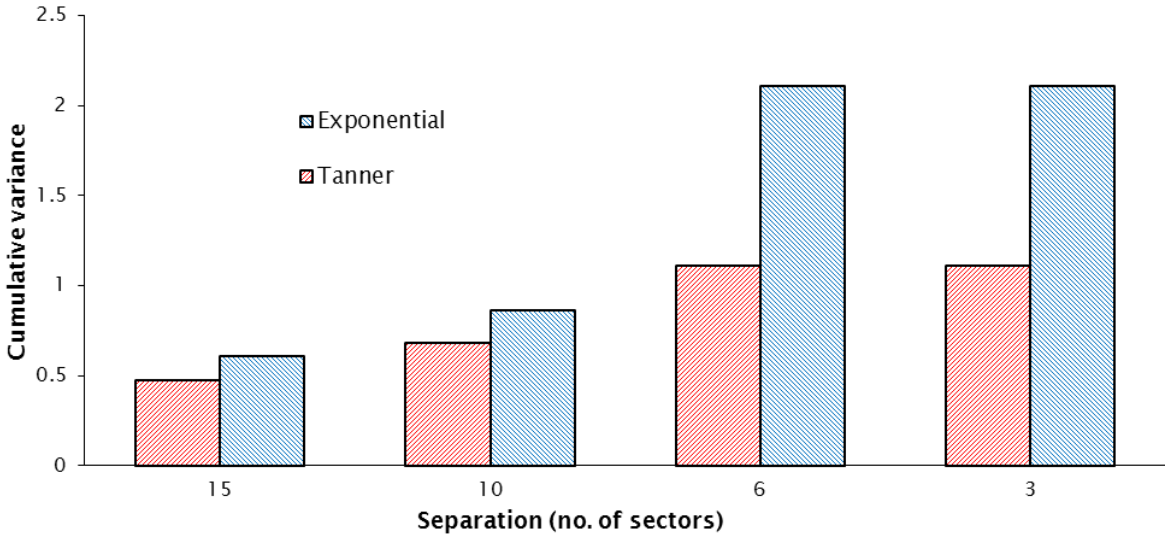


Figure 7.11 suggests that a 'sill' is reached with the aggregation into six sectors; no further effects are apparent with aggregation to three sectors. This may mark the range of correlation effects, but it seems large.

The cumulative variance for the Exponential deterrence function is larger for the Tanner function. This is probably due to the K factors compensating for a lack of fit in the Exponential function that is achieved in the Tanner. It is particularly marked at larger ranges, with aggregation to six and three sectors, beyond the range of possible correlation effects. It might be seen as an artefact of spatial correlation due to a lack of stationarity arising from an imperfectly specified systematic model.

It is difficult to be confident in these interpretations without results for the finer 65 sectors or 225 zones. It would also be helpful to have more points at large ranges, at aggregations to less than six sectors, to confirm the presence of a level sill; however, it is difficult to find any further aggregations of the six sectors in practice.

The horizontal axis of figure 7.11 is specified by the number of sectors into which the data is aggregated. This might be re-scaled to a continuous measure of separation (generalised minutes) using the average separations between segments shown in the lower half of tables 7.6 or 7.7. However, these show a narrow band of separations (68.38~73.32) which would not be greatly improved by disaggregation to the zonal level.

Table 7.9 Deterrence coefficients fitted with hierarchical K factors

Deterrence function	Term	Plain GLM		Hierarchical random K factors	
		Coefficient	SE	Coefficient	SE
Exponential	Cost λ	0.0638	0.0008	0.0743	0.0021
Tanner	Cost λ	0.0364	0.0013	0.0434	0.0026
	Log(Cost) γ	0.646	0.027	0.487	0.035

The fitted coefficients of the deterrence function are shown in table 7.9. The introduction of the random terms produces moderate changes in the coefficients; there is the usual trading-off between the correlated coefficients of the Tanner function. The standard errors all increase considerably, recognising the

uncertainty in the fit of the trip distribution which K factors imply. The cost deterrence effects are still highly significant.

This top-down approach is ultimately a crude approximation to a correlation structure, assuming that it is all captured within the segments specified. If all the error occurred at just one of the segmentations, as is implicit in a conventional application of K factors, it would appear as a fitted variance for that row of table 7.8 with zero for the other variances.

7.10 Bottom-up approach with correlation structure

The bottom-up approach is to fit a correlation structure based on separations. The fitting is again to an HGLM of the form

$$f(y) = X\beta + Z\upsilon$$

but in this case, there is a correlation between the random terms υ. In Genstat, the correlation matrix C is entered as a linearisation matrix L, which multiplied by its transpose gives the correlation matrix.

$$LL^t = C$$

The L matrix then post-multiplies the random factor Z, incorporating the correlation effects in a modified random vector Z*

$$Z^* = ZL$$

The treatment of correlated random effects in HGLMs is described by Lee et al (2006, chapter 8). Lip cancer rates in Scotland are analysed as an example of spatial correlation in section 8.6.2 and the program code is provided with Genstat. The correlation structures are based on the adjacency of administrative areas, rather than the continuous scale of travel cost used in this study.

Ideally the correlations would be formed between individual zone-to-zone movements, the bottom level for any aggregation to segments to which K factors might be applied. At this bottom level, the random factor Z would take a separate level for every zone-to-zone movement, the same as the individual units of the systematic regression. However, this greatly exceeded the limits of available computer memory. The random factor had to be set at a 10 x 10 segmentation to allow computation with correlation between the segments, losing shorter-range effects within the segments.

For the development of the HGLM computation, separations were based on WTSM HBW car costs, without removal of parking charges. The cost for each segment was a simple average of the component zone-to-zone costs. Separations were averaged by direction (PA+AP)/2 to give symmetry. Intrasector separations were set to zero to give unity on the leading diagonal of the correlation matrix. This is a coarse approach compared with regularisation undertaken later.

Correlations were calculated as a simple exponential function of separation

$$\rho = \exp(-\phi \times \text{Separation})$$

The L matrix was derived from the matrix C of correlations ρ by spectral decomposition using the Genstat directive `FLRV` to form latent roots and vectors

$$C = Q\Lambda Q^{-1}$$

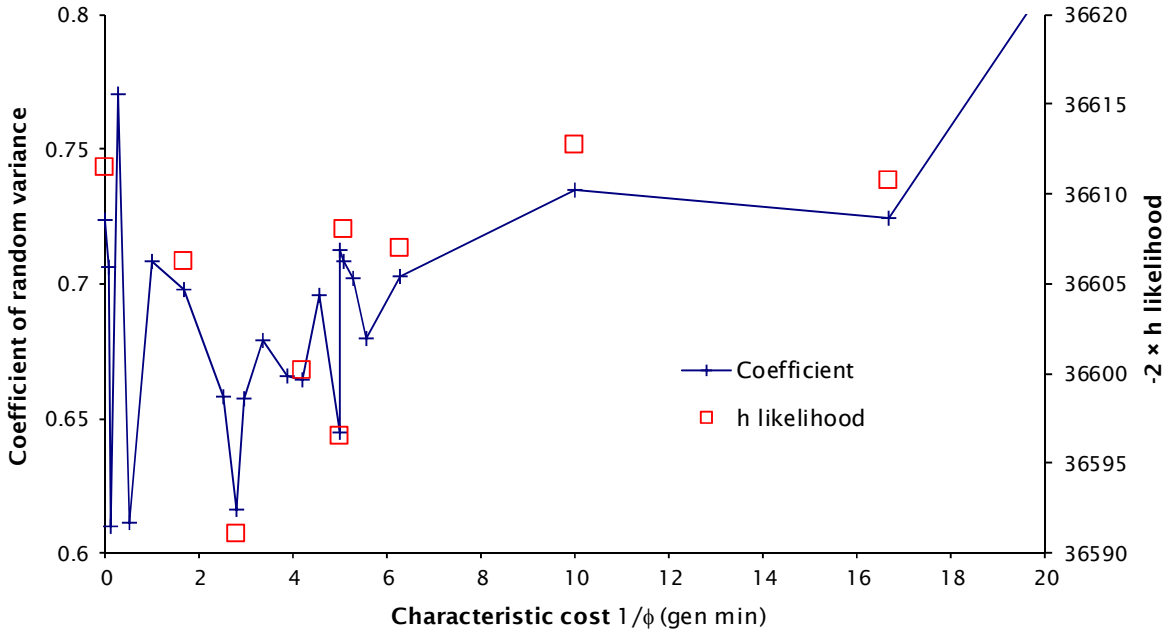
where Q is the square matrix of eigenvectors, which is orthonormal so $Q^{-1}=Q^t$, and Λ is the diagonal matrix of the eigenvalues.

Then $$L = Q\Lambda^{½}$$

giving the required equivalence $LL^t = Q\Lambda^{½}(Q\Lambda^{½})^t = Q\Lambda^{½}\Lambda^{½t}Q^t = Q\Lambda Q^{-1} = C$ since the transpose of the diagonal $\Lambda^{½}$ is itself, and the transpose of Q is its inverse

Genstat procedures cannot optimise parameters of the correlation such as ϕ while fitting HGLMs (Genstat's REML procedure for normal distributions can do so for many correlation structures). Therefore separate HGLMs were fitted with different correlation matrices from a range of values of ϕ, seeking an optimum in maximising the fit of the h-statistic, or minimising the variance of the random K factors fitted to the segments. The latter was much quicker to calculate than the former, and they show similar patterns. Both are plotted against the characteristic separation, $1/\phi$, in figure 7.12.

Figure 7.12 Variation in fit with spatial correlation



The vertical scale on the right is −2×(h-likelihood). This hierarchical likelihood is the recommended test for random effects (Lee et al 2006, section 6.5). It is a deviance, expected to approximate to a χ^2 distribution; in particular, a reduction by about 4 on the fitting of one parameter such as ϕ is just significant.

Both measures are very sensitive to small changes in the correlation, producing a noisy, jagged set of points rather than a smooth curve with a clear, credible minimum.

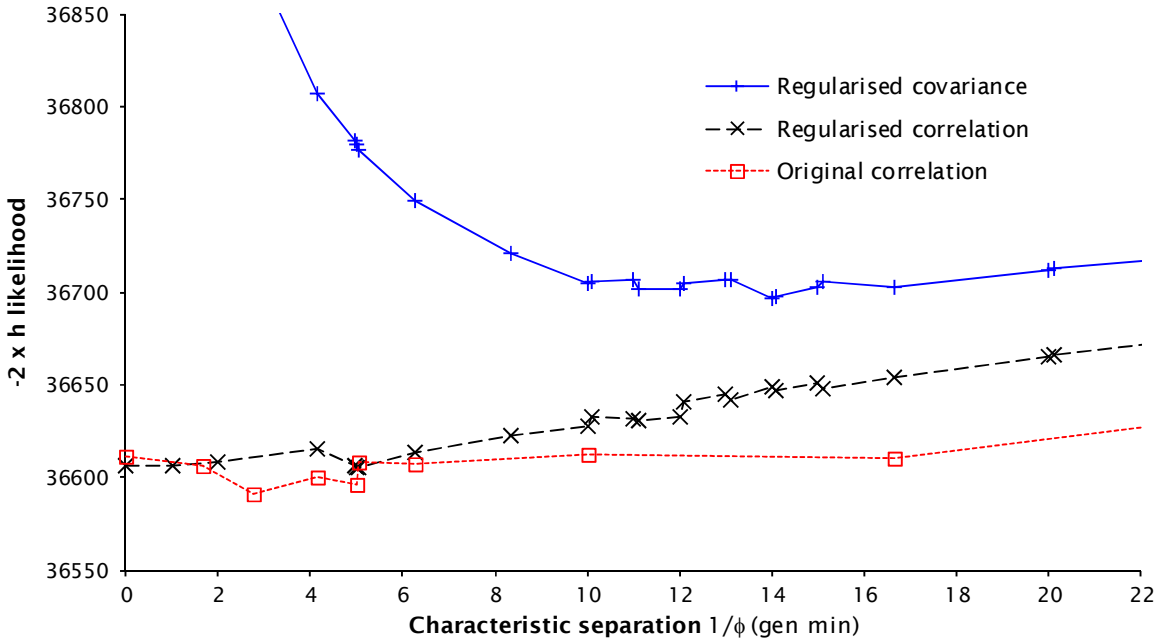
Although fitted K factors followed this noisy pattern, the totals of fitted trips appeared more stable. This suggests an interaction between the random K factors and the fixed trip end balancing factors. If the K factors were fixed, they would be aliased with the balancing factors; the structure may allow an incomplete convergence in a random model that has little practical consequence in terms of total trips, but leaves noise in the statistics of fit.

Similar perturbations of the measures of fit were found between versions 10 and 12 of Genstat, which were compiled with different compilers. Repeated runs with the same implementation of the software gave the same results.

At a later stage, models were fitted with covariance matrices produced by regularisation, such as those plotted in figure 7.9(c). These should incorporate the effects of heteroscedasticity and maintain the relationship between variances and covariances, unlike the original formulation of the correlation matrix.

Models were also fitted with correlation matrices normalised from the regularised covariance matrices, such as those plotted in figure 7.10(c). Their fit is shown in figure 7.13, together with the fit with the original correlation matrices.

Figure 7.13 Variation in fit with regularised covariance and correlation



The fit is shown just by h-likelihood, over a wider range. The fitted random variance has to compensate for changes in the variance included in the covariance matrices, as noted at the end of section 7.8.5. The covariance shows a much stronger effect of varying ϕ than the correlations, but still shows sensitivity to small perturbations around its broad minimum. Above all, even this minimum is a much worse fit than the correlations or even the simple models without any correlation, plotted on the vertical axis. (Their fits differ slightly because the costs used to fit the original still included parking charges.)

The systematic part of the model, $X\beta$, represented trip distribution with a Tanner deterrence function. The Tanner function reduced the risk of correlation in the random terms appearing from artefacts of an underspecified systematic model, but its two correlated terms, cost and log(cost), made their calibration harder to interpret, and may have made the model less stable.

Table 7.10 Deterrence coefficients fitted with correlated error terms

Coefficient of	Random terms		
	None (plain GLM) *	10 ×10 segments, uncorrelated	10 × 10 segments, correlated *
Cost, λ	0.0364	0.0440	c. 0.044 ± 0.0015
Log(cost), γ	0.638	0.4906.	c. 0.51 ± 0.01

* Costs include parking charges

However, table 7.10 shows that the fitted coefficients are not particularly sensitive to the different correlation structures. The major differences from the simple calibration arise mainly with the introduction of uncorrelated random K factors for 10 × 10 segments.

7.11 Computation

In developing these approaches, much effort went into overcoming computational constraints, particularly those of random access memory (RAM).

Increasing the installed RAM from 1 to 2 GB had no effect at all, due to limitations between Genstat and Windows XP. (There appeared to be a worthwhile benefit in speed.) 'Pinch points' in the code were identified where space problems arose. One was in the routine for higher-order Laplace transforms, which helps reduce bias in Poisson-normal models. This non-conjugate form had been retained for comparison with another algorithm, GLMM, but since this was not giving satisfactory results, the Poisson-normal form was abandoned in favour of the conjugate Poisson-gamma form. Another pinch-point arose in the calculation of h-statistics, which took 3+ hours compared with model fitting in c.5 minutes. Some apparently redundant code was omitted without noticeable ill-effect or marked improvement. It seemed unwise to go any further in altering the coding without a full understanding of it. The pinch points were reported to VSN, the developers of Genstat, who are discussing the computation of HGLMs with Youngjo Lee in Seoul.

VSN provided routines for fitting HGLMs with absorption by groups, which speeds up computation. However, only random variables can be so grouped, and trip distribution models include fixed factors with many levels for production and attraction zone balancing factors; grouping these could be more beneficial. The new routines can improve run times, but do not appear to avoid the pinch points noted above. They were supplied without source code, so the pinch points were identified in standard Genstat version 10 code. The grouping facility has been incorporated in later versions of Genstat.

Alternative approaches of migrating to a 64-bit Windows operating system or to the Canterbury University supercomputer were rejected as too demanding of the limited sources remaining. (A 64-bit OS was considered when specifying a new desktop PC a year previously, but offered no benefit then; approaches to use the supercomputer met with no response.)

Geographically weighted regression, developed by Fotheringham et al (2002), encounters similar problems in calculating exact measures of fit.

7.12 Alternative approaches

HGLMs comprise a very broad set of models with various fixed and random components. There are specific procedures for fitting some subsets of these models that present different features in their implementation, and offer alternative approaches to fitting K factors or other spatial error models. Some of these were tried to contrast with HGLMs, but were not pursued because of difficulties in comparison, and in order to concentrate on the problems presented by HGLMs.

Some fitting procedures in Genstat stem more from the analysis of variance (ANOVA) on categorical factors in designed experiments than from regression on continuous variates from unbalanced observations. Output formats differ and the correspondence of results is not always clear. There are also differences in the fitting of random terms with REML or its extension to GLMMs.

Fixed K factors can be fitted to finer segments by plain GLMs, but this requires more interpretation of their variance structure for comparison with h-statistics from HGLM. The fitted coefficients may be more sensitive to the effects of empty segments, since they are not expected to conform to a distribution like random terms. They may require adjustments to the degrees of freedom for empty segments.

Fixed K factors are used to fit within-segment distribution effects to complement the between-segment effects analysed in the next chapter in section 8.7.3.10.

Estimates can be found for both fixed and random K factors. However, the fixed effects are expressed in GLMs as differences from a base segment, while the random effects are centred about a zero mean in HGLMs. With differing trip end balancing factors as well, comparison is difficult.

A negative binomial model could be fitted to represent a bottom-level Poisson-gamma error structure with one random K factor for every zone-to-zone pair. However, this would omit the quasi-Poisson allowance for other errors made by setting DISP=* in other models. Alternative parameterisations of the gamma distribution can make comparison of results from with other models awkward.

In Genstat, the weights for a GLM can be specified as a matrix, rather than a simple variate. This could be used to try different correlation structures, the bottom-up approach. However, the structures would still be very large at the bottom zone-to-zone level and could not be applied at intermediate levels without aggregating all the data. Covariance matrices would need to be inverted to act as the weight matrix. The distribution of random error implicit in this approach is not defined where the main distribution is not normal and the approach is not commended by VSN.

The REML procedure is restricted to normal errors for both main and random terms. It can find the best parameters for certain correlation structures within its fitting process, which could remove the need for the trial-and-error fitting of ϕ in HGLM or GLM.

The GLMM procedure is an extension of REML which allows non-normal distribution of the overall error, but is still restricted to normal distribution of the random terms. There is no provision for weighting, so weights cannot be used to adjust for survey expansion. With the simple, single overall weight used in this exercise, this could probably be overcome by judicious scaling and manual adjustment, but would further complicate analysis and comparison.

One facility available in a REML model and GLMM but not in HGLM is the inclusion of continuous variates in the random term. Specifying the random term as, say, Segment.cost fits a random variation in the cost coefficient between segments. In mixed modelling terms, this can be interpreted as a random coefficient (Train 2003, section 6.2). In gravity modelling terms it is a random L factor, extending the random model from the K factor, which represents a constant in the deterrence function, to the coefficient of cost.

A thorough analysis of random L factors needs to establish its main effects among production or attraction zones (or sectors, households, persons or trips) before investigating its interaction effects among zone-to-zone movements or segments. The main effects of K factors are absorbed in the trip end balancing factors.

Emmerson (2008) showed improvements in trip distribution models for Strathclyde and Leeds when cost coefficients for Exponential or Power deterrence functions were allowed to vary by zone. Variation by destination zone tended to fit better than by origin, though not by much. These ‘zonal’ models fitted fixed L factors by GLM.

It is possible that random L coefficients can be modelled in HGLM by introducing the cost term via the structured dispersion model (Lee et al 2006, chapter 7; `DOFFSET` or similar parameter in the `HGRANDOMMODEL` procedure of Genstat).

Thinplates are two-dimensional splines offering a smooth surface. They might be used for an empirical investigation of spatial variation in trip distribution similar to the fitting of ordinary splines as deterrence functions in section 4.5. Since a single surface could only represent one end of a trip, this would be

limited to the main effects (ie determined at production *or* attraction end) of L factors in the first instance. This approach would be closer to the geographic weighted regression of Fotheringham et al (2002).

7.13 Lorelogram

In conventional spatial statistics, the spacing of correlations is explored in variograms formed from direct observations of normally distributed random variables. The very sparse Poisson residuals from trip distribution do not correspond well with these. Diggle (pers comm) suggested use of the lorelogram, an examination of the log odds ratio for binary data developed by Heagerty and Zeger (1998).

7.13.1 Basis

In binary data there are two possible outcomes – failure or success, absence or presence – denoted here as Y = 0 or 1. Correlations between pairs of outcomes Y_1 and Y_2 are investigated by calculating the odds ratio

$$P[Y_1 = 0,Y_2 = 0] \times P[Y_1 = 1,Y_2 = 1] \; / \; P[Y_1 = 0,Y_2 = 1] \times P[Y_1 = 1,Y_2 = 0]$$

If Y_1 and Y_2 are independent, the joint probabilities comprising the odds ratio are simple functions of the individual probabilities P_1 and P_2, eg

$$P[Y_1 = 1,Y_2 = 1] = \quad P_1 \times P_2$$

$$\text{and } P[Y_1 = 0,Y_2 = 0] = (1-P_1) \times (1-P_2)$$

The odds ratio then reduces to unity. Positive or negative correlations between Y_1 and Y_2 give odds ratios more or less than unity. A plot against separation shows a simple spatial correlation as large odds ratios at short separations decaying to unity at larger separations.

An experimental odds ratio for observations can be found simply from counts of paired outcomes into a 2 x 2 contingency table.

Table 7.11 2 × 2 contingency table

	$Y_2 = 0$	$Y_2 = 1$	Sum
$Y_1 = 0$	N_{00}	N_{01}	N_{0*}
$Y_1 = 1$	N_{10}	N_{11}	N_{1*}
Sum	N_{*0}	N_{*1}	N_{**}

The odds ratio is then $N_{00} \times N_{11} \; / \; N_{01} \times N_{10}$

These simple tabulations can be computed at the level of zone-to-zone movements without great difficulty. The tabulations are also stratified by separation to form a lorelogram.

7.13.2 Preparation

The observed trip matrix was reduced to binary data by setting cells with any observed trips to 1, leaving all other cells as 0. This loses information from matrix cells with more than one observation. About half of the WTSM observations fall in such cells, even in the land-use home-workplace formulation that omits return trips.

The lorelogram was formed from all pairings of movements between zones with observed trip ends, excluding self-pairings but including pairings with only production *or* attraction zone in common.

7.13.3 Systematic effects of trip distribution

The simple tabulation approach that permits computation at the level of zone-to-zone movements does not offer any allowance for systematic variation in the marginal probabilities P_1 and P_2, which is to be expected from trip distribution.

Heagerty and Zeger (1998, section 3.1 and appendix A) propose joint modelling of systematic and random effects. This appears similar to the HGLM approach, but with general estimating equations (GEEs) rather than GLMs. Since the expectation for every cell in the trip matrix is different, this posed a similar problem of modelling a large number of individual units, but without established software.

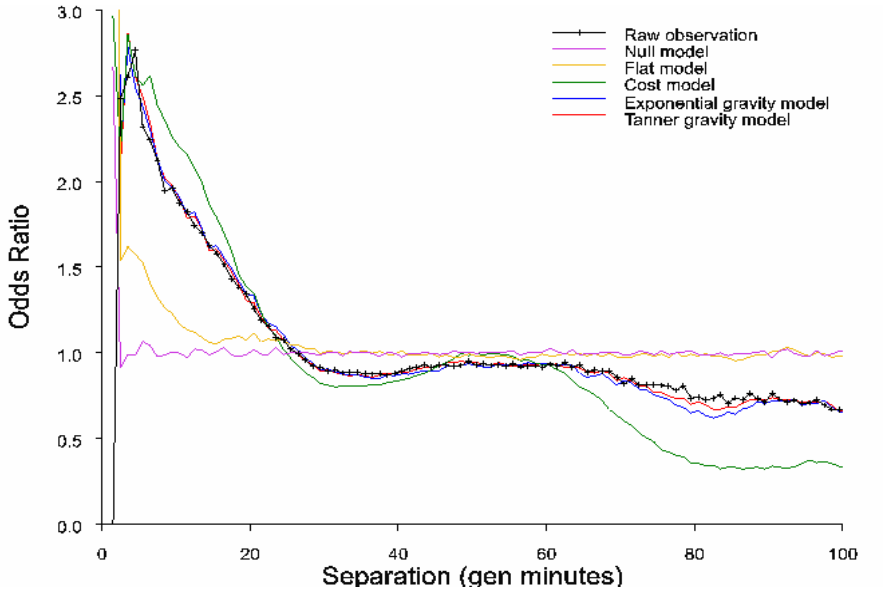
Lacking a simple allowance for the systematic effects of the gravity model, a lorelogram was formed from the raw observations. As shown in figure 7.14 this takes the form expected from spatial correlation, with marked increases in the odds ratios as separations decrease below 20 generalised minutes. The odds ratios also fall below unity for some larger separations.

7.13.3.1 Control models

As an empirical check on the process, lorelograms were also formed from trip matrices that had been synthesised from models and then randomised. All cells within each matrix were randomised independently, so there is no spatial correlation in the error structure. The randomisation was by a Poisson distribution about the modelled cell means, representing the sampling of observations. The means always sum to the same as the raw observations over the whole matrices. The randomisation process is consistent across models in that cell A is always at, say, the 35%ile, cell B always at, say, the 76%ile etc of the Poisson distribution about its modelled mean.

Since none of modelled data has spatially correlated errors, their lorelograms were expected to be flat at unity. The lorelograms are shown in figure 7.14 and the systematic models used to generate them are described in the following sections.

Figure 7.14 Lorelograms



7.13.3.1.1 Raw observation

Home-workplace pairs with car trips in the Wellington HIS. This is the data in which spatial correlation is sought. An effect does appear in the lorelogram as an odds ratio of c.2.5 for small separations. However, this should not appear in any of the following models.

7.13.3.1.2 Null model

This model has the same mean for all cells. It is the only model that shows an absence of correlation in having an odds ratio of unity at all separations. At the very shortest separations there are very few pairings so random noise appears.

7.13.3.1.3 Flat model

Cell means are proportional to the production and attraction trip ends of the observations. This is the usual base for trip distribution, allowing for the spatial distribution of homes and of workplaces, but not for any deterrence effect of the cost of travel between them. There is a small correlation effect for small separations, less than for the observations, but nonetheless spurious when looking for correlations in model residuals.

7.13.3.1.4 Cost model

In contrast with the flat model, cell means are scaled only by the deterrence effect of cost (for an Exponential function calibrated to observations) and not by trip end generations or balancing factors. The apparent correlations are more marked than for observations and show the same dips below unity for greater separations.

7.13.3.1.5 Exponential gravity model

A conventional trip distribution model, with trip ends balanced to observation and cost deterrence function calibrated to observations. The Exponential form is consistent with basic choice modelling theory.

7.13.3.1.6 Tanner gravity model

An extension of the Exponential distribution function (also known as gamma or two-parameter) which generally gives a better fit to the Wellington data.

Since there is no spatial correlation between the errors in the modelled datasets, the increased odds ratios for small separations must be artefacts of the systematic spatial distributions, ie the gravity models. These artefacts are apparent even in the sub-components of trip end generations and cost deterrences.

The lorelogram for the raw observations corresponds closely with those of the gravity models. It is entirely plausible that there is no spatial correlation in the observed data beyond the established systematic effect of gravity models. By the same token, lorelograms give no indication of anything but a close fit of the gravity models to the observations.

7.13.4 Stratification

Attempts were made to reduce or remove the appearance of spatial correlation due to the systematic effects of trip distribution by stratification. Stratification is a relatively simple extension of the tabulation process used to accumulate contingency tables by separation band.

The stratifications tried were by the difference in travel costs or the ratio of probabilities (from gravity models) between the two movements in a pair. For example, a separate lorelogram would be formed for, say, pairs of movements whose travel costs differed by between 5 and 10 generalised minutes. The stratifications generated a wide variety of lorelograms, but they follow the pattern of figure 7.14 in that:

- the plots for the gravity models are similar to those for observations
- only the null model shows no apparent spatial effect.

Thus such stratification seemed not to offer a way of removing the systematic effects of trip distribution from the lorelogram on raw observations.

From later empirical exploration of mixed populations (below), odds ratios of unity were obtained from combining non-correlated populations with either P_1 or P_2 in common. This suggests that a stratification by the probability of observing a trip in just one of the movements in each pairing may be more useful.

In the limit, a stratum would comprise pairings with just one movement, but since the outcome for this movement (observed trips or none) would be the same for all pairings, one row or column of the contingency table would all be zero and the odds ratio would have no meaning.

Stratification still leaves the problem of aggregating or interpreting information across the strata.

7.13.5 Odds ratios from mixed contingency tables

The appearance of spatial correlation in the lorelograms seems to be an artefact of trip distribution which produces different probabilities P_1 and P_2 in each pairing. The effect was explored empirically by generating contingency tables made from a simple mixture of two populations. Each population had consistent values of P_1 and P_2, but these differed between the two populations. There was no correlation between outcomes Y_1 and Y_2 in either population. Different sizes of population N were considered.

An example of the contingency tables expected from such populations, individually and in combination, is shown in table 7.12.

Table 7.12 Contingency tables from mixed populations

a) First population: N = 100, P_1 = 0.2, P_2 = 0.2

	$Y_2 = 0$	$Y_2 = 1$	Sum
$Y_1 = 0$	64	16	80
$Y_1 = 1$	16	4	20
Sum	80	20	100

Odds ratio = 64 x 4 / 16 x 16 = 1

b) Second population: N = 100, P_1 = 0.5, P_2 = 0.5

	$Y_2 = .0$	$Y_2 = .1$	Sum
$Y_1 = 0$	25	25	50
$Y_1 = 1$	25	25	50
Sum	50	50	100

Odds ratio = 25 x 25 / 25 x 25 = 1

c) Combined populations: adding the contingency tables

	$Y_2 = 0$	$Y_2 = 1$	Sum
$Y_1 = 0$	89	41	130
$Y_1 = 1$	41	29	70
Sum	130	70	200

Odds ratio = 89 x 29 / 41 x 41 = 1.54

Simple mixtures can give odds ratios less than unity, or much greater than unity. Odds ratios equal to unity occur when there is a common value of either P_1 or P_2 in both populations.

Exchanging P_1 with P_2 between the two populations, as will occur by including both permutations of a pairing, does not generally lead to an odds ratio of unity.

The populations which are mixed will not usually be independent when forming the contingency table for a particular separation in a lorelogram. Each pairing that is counted into a contingency table is made up of two production–attraction movements; the same movement will appear in many pairings; and the pairings that fall into a particular separation band will be constrained by spatial geometry. Geometric constraints may be stronger at small separations where:

	Pair A is close to Pair B
and	Pair B is close to Pair C
implies	Pair A is close to Pair C.

These constraints of pairing and geometry are akin to those on authorised theoretical variograms and correlation structures.

7.13.5.1 Universal odds ratios

The accumulation of all pairings (including self-pairings) within any binary dataset will always give an odds ratio of unity. If a real spatial correlation produces odds ratios of greater than zero for short separations, this may place a constraint on the longer separations where the odds ratio could be expected to tend to unity. The self-pairings are intrinsically correlated.

7.13.5.2 Simpson's paradox

The effect of an association appearing in a combined population where none exists in sub-populations may be seen as an aspect of Simpson's paradox. Simpson's paradox is usually described in terms of a reversal in the sense of association between an independent variable X and a dependent variable Y, rather than between two outcomes Y_1 and Y_2. The existence of differing sub-populations is described as a lurking variable or a third dimension Z of the contingency table.

It limits the lorelogram as a simple graphical tool for exploring correlation in residual errors to binary data with constant probabilities, and requires similar homogeneity for the general application of contingency tables, one of the most basic and common statistical methods for detecting association.

It emphasises the need to identify and ensure the appropriate stationarity conditions when undertaking geospatial analyses.

It offers a plausible mechanism for the appearance of spatial structure in lorelograms formed from gravity models synthesised and randomised without any such correlation in the random model.

7.14 Possible applications

Since this study has failed to demonstrate either a practical method of geospatial analysis or evidence of spatial correlation in the data that calls for one, this consideration of a practical application is speculative.

Geospatial analysis has been approached here as an analytical tool to examine a possible structure for errors in trip distribution modelling. Although several possible mechanisms leading to spatial correlation are put forward in section 7.3, most suggest other models related more directly to the mechanism. If spatial correlation were detected, it might be treated as an error stratum within which to test the explanatory power of these mechanisms. Alternatively, the concept of space might be extended from geographic space, or that defined by the transport network, to socio-economic space, considering similarities and separations in household and employment characteristics.

Current patterns of travel may reflect historic patterns of land-use and transport networks. This is often reflected in the lagged variables of land-use models. Correlated error terms can also fit time series, but consistent series of trip distribution data are not common.

A clear knowledge of spatial correlations in trip patterns could inform the design of travel surveys. While it might not change what is regarded as good practice, it could clarify the benefits of costly dispersion in sampling and allow a more economical design.

7.14.1 Model building with K factors

At an empirical level, K factors are used to adjust for mismatch to observations in trip distribution models.

The top-down approach provides K factors which can be carried forward into a future year model. Their values will be less extreme than K factors fitted as fixed effects, because as random terms they are also constrained to fit an error distribution. This is the 'shrinkage' effect of mixed modelling.

If the bottom-up approach could be applied at the preferred zone-to-zone level, it would produce a zone-to-zone matrix of K factors. This would be similar to a matrix for 'adjusting' a demand model to an observed matrix. However, it would provide a rational smoothing between nearby movements and may take useful (non-zero) values where there are no trips observed for a cell.

Basic geospatial models provide smooth predictions across observed points only if there is a nugget variance. Smoothing might need measures to ensure a nugget variance such as non-zero intrazonal separations in an exponential correlation function, but the overall Poisson distribution of the sampling process may also prevent estimates simply equalling observations where they occur.

7.14.2 Zone prediction for land-use development

The trip distribution from a new zone or development can be predicted from the top-down approach by identifying the sector in which it lies and applying the appropriate K factors.

Formal prediction from the bottom-up approach could be by a weighted averaging of nearby K factors, using the methods of kriging to find the optimal weights.

Both are similar to the simple practical approach of copying the trip distribution of a new development from a similar existing development in the vicinity. However, it is the K factor adjustments to the trip distribution that is interpolated from the local sector or area, rather than the trip distribution as a whole.

Both depend on the credibility of the spatial mechanism for determining variations in trip distribution represented by K factors. If this is not accepted, new development might be treated as a 'new draw' from the random error distribution, or lying beyond the range of the correlation function from any existing development. There would be no K factor for the systematic model, but the error of prediction would be greater.

This might be applied to predictions into the far future to represent the decay of existing patterns with increasing uncertainty.

7.14.3 Scheme prediction for network development

The movements served by a new link or public transport service will tend to have similar origins and destinations and thus are liable to be correlated if spatial correlation exists. Taking this correlation into consideration may show that the predictions of usage or benefits are more accurate than if errors in estimating the movements are assumed to be independent. This contrast would require careful formulation of the equivalent random error or overdispersion in the null case without spatial correlation.

The set of movements using the new link or service can be seen as an irregular block for kriging.

Predicting the accuracy of a total of many estimates from a simpler GLM can be computationally demanding and this level of consideration of the accuracy of a model output is far beyond usual practice.

7.15 Summary

The analyses have produced no firm evidence of a spatial structure in the residual errors of the fitted gravity models.

The top-down approach suggests that there could be a structure with quite a large range, ending only at aggregation to the six sector level. However, it is hard to be confident in this interpretation in the absence of 65 sector and zonal strata from the analysis, and recognising from regularisation that the model may be poorly specified with all-or-nothing correlations in the block structures and without accounting for heteroscedasticity induced by irregular sectors.

The clearest effect in the bottom-up approach appears to be that of the heteroscedasticity incorporated in the regularised covariances. This produces a worse fit than simply fitting K factors to 10×10 segments with no correlation between them.

Lorelograms for modelled data without spatial correlation in their residuals are very similar to those for the raw observed data, offering no evidence of spatial correlation in the observations. However, given the already strong appearance of spatial effects in the lorelograms, presumably due to the systematic effects of trip distribution, it is not certain that a spatial correlation of residuals would be discernable.

Correlations between zone-to-zone movements can involve large datasets. If the number of zones is N_{zone}, the number of zone-to-zone movements is N_{zone}^2, and the number of pairings of movements which may be correlated is N_{zone}^4 (including intrazonal movements and self-pairings). The WTSM internal zoning system is relatively modest by current standards with 225 zones, but this gives 2,562,890,625 pairings of movements. This is a substantial number for computer systems in common current use (2007–2009) and even the fairly simple calculations of regularisation and raw lorelograms require some care when working on this scale. The more complex mathematical processes involved in fitting statistical models on such scales can easily exceed the power of ordinary personal computers.

The exercises in regularisation demonstrate that, if the range of spatial correlation is on the same scale as zone size, aggregation of observations from zones to sectors will lose much of the information about the spatial correlation.

The regularisation also shows that, even if the spatial error process is simple and regular at the punctual or zonal level, it will become irregular with heteroscedasticity under aggregation to irregular zones or sectors, which are usual in practical transport models.

This heteroscedasticity may provide useful information about the spatial error process. Neither of the analytical models took advantage of this, or made allowance for it save for a few late bottom-up models with covariance matrices produced by regularisation.

At present, HGLMs and their application to geospatial modelling are too demanding and immature for practical application to full trip distribution. The calculation of h-statistics appears to be as computationally demanding as fitting HGLMs themselves. Compared with the simpler log-likelihood deviance measures used to assess the fit of ordinary GLMs, the properties of h-statistics are not so well established; there is not the same depth of experience or provenance of practical application.

Experience with the lorelogram and with the larger K-factor variances for the Exponential model in figure 7.11 demonstrates that systematic effects of trip distribution can all too easily appear as random spatial effects. The importance of stationarity as a condition for so much geospatial theory suggests the fit of spatial error models is always liable to be sensitive to the specification and fit of the systematic trip distribution model.

A better approach to the use of geospatial statistics in transport modelling would be to develop them from smaller, simpler problems, possibly working with artificial or synthesised datasets of known properties. However, generating random samples with a spatial error structure is a sophisticated process in its own right; larger zones raise issues of MAUP; and smaller study areas raise issues of ergodicity.

A sensible simplification would be to start with trip end generation models, rather than trip distribution models. This will reduce the complexity of the problem; correlation matrices will be of the size N_{zone}^2 rather than N_{zone}^4. The next chapter shows small adjustments to trip ends allow a synthetic trip distribution model to meet validation criteria for assigned flows; these adjustments have strong spatial patterns. It is also good statistical practice to investigate main effects, such as trip generation, before their interactions, such as trip distribution.

8 Model estimation from aggregate data

8.1 Introduction

Analysis in previous chapters has required trip information to be identified with specific pairings of production and attraction zones. The example has been an observed trip matrix built from the Wellington household interview survey (HIS). Such production–attraction (PA) information is generally expensive and inconvenient, requiring interviews with travellers.

On the other hand, the volumes on links in the network are relatively easy and cheap to count. The information may already be available from sensors for traffic signals or ticket sales on public transport. However, these counts are usually aggregates of many different PA movements which are not readily dissociated using the count data alone.

The generalised linear models (GLMs) used so far do not accept aggregate data. However, the availability and cheapness of aggregate count data are exploited in transport modelling by the techniques of matrix estimation.

Matrix estimation seeks to create or adjust a matrix so that, when assigned to a network, the modelled flows match observed counts. Conforming to these observations leads to an ad hoc model structure which is determined by the availability of data. A typical form is

$$T_{ij} = t_{ij} \prod_k X_k^{R_{ijk}}$$

where the initial matrix t_{ij} is multiplied by a set of factors X_k. Each factor X_k corresponds to a count k and applies to the $_{ij}$ movements included in that count, indicated by R_{ijk}. This may be solved simply by iterative factoring akin to Furness, and can provide an exact fit to a consistent set of counts.

Willumsen showed that this form is a maximum entropy solution with the counts as fixed constraints. Cascetta (1984) and Spiess (1987) allowed for errors in the counts as well as in the initial matrix, and found maximum likelihood solutions for normal and Poisson errors respectively. Bell (1984) demonstrated the similarities between these approaches.

There is no inherent cost deterrence function but one can be introduced as the initial matrix, eg

$$t_{ij} = c_{ij}^{-\gamma} \exp(-\lambda c_{ij})$$

where c_{ij} is the cost. To calibrate a trip distribution, the cost coefficients γ and λ have to be optimised.

One matrix estimation program, MVESTM, offers this capability. Its adjustment of the initial matrix takes the basic form

$$T_{ij} = t_{ij}\, a(i)\, b(j) \prod_k X_k^{R_{ijk}}$$

where a(i) and b(j) are row and column factors that can act as the trip end balancing factors of a trip distribution. The count factors X_k can be suppressed; since they can fit a consistent set of counts without a cost model, they will absorb the effects of cost deterrence if allowed to remain. The resulting model

$$T_{ij} = a(i)\, b(j)\, c_{ij}^{-\gamma} \exp(-\lambda c_{ij})$$

is the form of a trip distribution with a Tanner deterrence function.

Fitting this form of model to aggregate data has been explored with MVESTM, which maximises the likelihood of Poisson errors and so is consistent with previous calibration by GLM. The suppliers of MVESTM, Citilabs, kindly provided source code for the shell program and a description of the intercept file structure, allowing considerable manipulation and modification, together with the Cube transport

modelling suite in which it is now known as Cube Analyst. MVESTM's development, key capabilities and application are set out in section 8.2 onwards. Further details are given in appendix E. Alternative computational approaches are discussed in section 8.8.3.

In practice, matrix estimation is usually employed to find an empirical base year matrix that fits well with observed counts and other data. The structure of the model follows that of the data available. The resulting matrix can be used for assignment in a traffic model, or as a base for forecasting with a demand model, but this form of matrix estimation cannot generate future year matrices directly.

By contrast, trip distribution calibration seeks parameters for a theoretically defined model, such that it matches observations. The parameters are carried forward to model travel demand with different land uses and networks. This approach is referred to here as 'model estimation'; its contrasts with conventional matrix estimation and with matrix building are discussed further in section 8.8.5.

Model fit has been interpreted statistically. This required the proper scaling of the uncertainty or errors in the input data. Validation criteria for transport models have been considered in setting the scale of discrepancy expected for screenline counts.

As a practical example, trip distributions were calibrated from counts at four screenlines in the Wellington region, without recourse to a fully observed PA matrix. As with previous calibrations by GLM, the WTSM provided the data and framework.

The practical problem of calibrating an all-day model of trip distribution between productions (homes) and attractions (workplaces) from counts of traffic by direction and period was addressed by incorporating period and direction factors in the intercept proportions. The manipulation is described in section 8.6.3.3.

The analysis was confined to a single purpose and mode – commuting by car. Traffic counts were factored down to this purpose using the WTSM. The extension of the methodology to multiple purposes is discussed in section 8.8.1.3.

Analysis starts in section 8.7 with the fully observed PA matrix from the HIS. This had been calibrated by GLM, so the consistency of the two approaches could be demonstrated.

Computational limitations required aggregation of the HIS matrix. Knowing the full zonal matrix, this gave useful insights into aggregate data. It also provided a step-by-step progression, via single segments and quadrants, to a calibration broadly comparable with that from a two-way, all day count at one of the actual screenlines.

Counts at the four screenlines in three periods and two directions are analysed in section 8.7.7. Combinations are introduced either as separate items of information, or aggregated into a single count. The same data is presented in either case, but in a different form.

The analyses showed that calibration in Wellington needs trip end information as well as screenline counts. Theoretical consideration of the information needed for trip distribution modelling in section 8.5 suggests that this is generally the case in practical studies. However, planning data (eg population and number of workplaces) might provide the pattern of trip ends for a joint calibration of generation and distribution.

The sensitivity of calibration to trip end information was explored with strong and weak trip end confidences. The corresponding latitude in fitted trip ends affects a model's ability to meet validation criteria for fit at screenlines. Section 8.7.7.6 shows that, for the same data and model structure, a synthetic demand model fails validation if conventionally distributed from fixed trip ends, whereas a matrix estimated without trip end constraints passes with ease.

8.1.1 Nomenclature

To describe aggregation, **sectors** refer to groupings of zones – a part of the study area.

Segments refer to groupings of PA or OD movements – a part of the matrix. Segments are often defined by the intersection of a production sector and an attraction sector; sometimes by the movements intercepted at a screenline; but not usually a set of complete rows or columns in the matrix, which would be a sector.

Productions and **attractions** are the home and work end of a trip; **origins** and **destinations** are the start and end. In general, they are the same in the morning when commuters go to work, O>D=P>A, and reversed in the evening when they return home, O>D=A>P. Calibrations here are of production–attraction (**PA**) matrices, related to land uses; the direction of counts is determined by origin and destination (**OD**).

All full-scale analysis is by MVESTM, and some of its terminology is adopted for this chapter. It expects to accept counts across **screenlines** rather than on individual links. This term is adopted for any aggregation of movements for which a total of trips is presented to MVESTM as a screenline count. This can be a traffic count at an actual screenline, the total for a segment of an observed trip matrix, or even the trips from a single cell of the matrix.

One of the four actual screenlines taken for the example forms a cordon around the centre of Wellington; another comprises a single counted link. All four are referred to as screenlines. The direction of traffic crossing these screenlines is described as inbound towards the centre of Wellington, and outbound away from the centre. The three counting periods are: AM, 7am–9am; IP, (interpeak) 9am–4pm; and PM, 4pm–6pm.

Cost parameters are named in MVESTM as alpha and beta in the function $c_{ij}^{alpha} \exp(-beta \times c_{ij})$ where c_{ij} is the cost. As single parameters for Power or Exponential deterrence functions respectively, alpha will be negative or beta will be positive for the usual decrease in travel with cost. Alpha and beta correspond with the coefficients $-\gamma$ and λ, which are the symbols used in this chapter as in the rest of the thesis.

The uncertainty or error in data input to MVESTM is termed confidence. It is not the same as a statistical confidence interval. In statistical terminology it is inversely proportional to the index of dispersion, which is the variance divided by the mean. Confidence is equivalent to weighting in previous chapters; in MVESTM's technical documentation the symbol λ is used, but the symbol 'w' for weight is adopted here. The confidences input to the program are scaled up by 100, ie confidence = 100 $\times$ weight or 100/(index of dispersion).

8.1.2 Model estimation from counts

Although recent practice has concentrated on the estimation of empirical base matrices from counts, there is a long history of methods for calibrating models from counts.

Low (1972) generated uncalibrated synthetic matrices of 'trip probability factors' of the form $f = A_i P_j t_{ij}^m$, where A and P are measures of production and attraction such as population and employment. Matrices for n different purposes were assigned to the network and the assigned volumes were compared with observed ones.

$$V_{observed} = a + b_1F_1 + b_2F_2 + \ldots + b_nF_n$$

The coefficients b were estimated by multiple regression, allowing unobserved, 'scheme' or future year link flows to be forecast by the same procedure with appropriate assigned networks and generation measures P_i and A_j. However, the coefficients b are essentially trip generation rates, and no method was suggested for calibrating the trip deterrence function t_{ij}^m. An inverse square of time was used in an example.

Jensen and Nielsen (1973) took a broadly similar approach, but with a single purpose, and population as the measure of generation. As well as fitting a rate to this measure, the individual 'potential' for generation by each zone was estimated where there were more counts than zones. This improved the

model, but the potentials of adjacent zones were sometimes hard to distinguish and the correlation between them was calculated. Cost deterrences were calculated from journey times at each iteration of a congested assignment and averaged in parallel with the volumes. Different coefficients for a power function were tried and an inverse power of 2.25 fitted best.

Fitting was by least-squared errors normalised by observed counts. A later paper (Holm et al 1976) found that this Poisson-like model for variance proportional to the mean was a better description of stochastic errors than a constant variance or a standard deviation proportional to the mean. An inverse power of 3.25 was found for the trip distribution function. Accuracies of traffic volumes modelled in a rural area of South Zealand were found to be as good as those from conventional Danish traffic models.

Overgaard (1972; in OECD 1974) drew on information from Aarhus in formulating a generation model for Silkeborg, distinguishing apartments from single family houses. Trip rates were calibrated by counting on cordons around two zones, one of each housing type. Different power laws were tried for trip distribution models; the crossings of three screenlines were compared with observations. The best fitting function increased linearly with travel time up to 1.5 minutes and then diminished with an inverse power of 1.8.

Carey et al (1981) considered the estimation of direct demand functions, similar to those above, as optimisations of quadratic programming problems arising from least-square minimisation.

Robillard (1975) used a gravity formulation $T_{mn}=R_mS_nG_{mn}$ to resolve individual OD movements which were otherwise not unique in sets of linear equations from link volumes. Although the quantity G_{mn} was described as measuring the relative cost of travelling from origin m to the destination n, it played the role of deterrence function rather than cost, and no attempt was made to relate the two.

Hogberg (1976) recovered five parameters from a synthesised generation and distribution model. There were three purposes, each with one trip rate parameter, and two parameters for a deterrence function. All three purposes were singly constrained with the same deterrence function, a quadratic in the logarithm of cost after Bexelius. Randomisation was applied to synthesised OD movements rather than to assigned link volumes. Calibration was by minimising least-square differences of volumes on half the links; the other half were retained to demonstrate the fit.

Wills (1986) compared intervening opportunity and gravity models for trip distribution, using the Box-Cox transformation to combine the two into a flexible gravity-opportunity model. He drew on his unpublished postgraduate studies into fitting trip distributions from count data. As a case study, distributions were fitted to counts on 112 highway links between 58 communities in Ontario, Canada. Their trip generations were modelled as powers of their populations; three separate parameters were fitted for productions, attractions and intervening opportunities. A single parameter was fitted to a Power deterrence function of distance; this was close to -2, the inverse square of the classical gravity model. The model was fitted to give least-squared errors in link volumes. Their likelihood ratio showed a statistically significant improvement over a pure gravity model. Double, single and unconstrained (direct demand) distribution models were considered, with all-or-nothing assignment.

Tamin and Willumsen (1989) considered alternative assignment methods and objective functions in their application of gravity-opportunity models to Ripon, UK. There were 63 counts on 188 one-way links between 26 zones, seven of them external. Trip generations were taken from a transport model. Assignment was all-or-nothing, or Burrell at two levels of randomisation. Fitting was to least-squares, weighted least squares, or maximum likelihood. Least-squares corresponded with a normal error in counts, while the weighting by the inverse of the observed volume and maximum likelihood of a multinomial distribution of counts suggested Poisson-like errors. No formal statistical tests were made on these objective functions of the fit to link counts; the main comparisons were between the estimated

matrices and a matrix observed from roadside interviews. Even the best distribution models showed a relatively small improvement over a Furness model, presumably a simple proportioning by trip ends without gravity or intervening opportunity effects. This may be due to the small study area limiting deterrence effects, or sampling noise in the observed matrix.

Tamin and colleagues at the Institute of Technology Bandung in Indonesia have published several similar papers on this 'unconventional methodology'. Tamin et al (2003) estimated a joint mode-choice and trip distribution from passenger counts on public transport. Exponential, Power and Tanner deterrence functions were fitted, with some unusual coefficients. Suyuti et al (2005) reported the estimation of gravity models from link counts with Bayesian inference and maximum entropy objective functions as well as least squares and maximum likelihood. The coefficients found for the Exponential deterrence function differed considerably. An equilibrium assignment gave a markedly better fit than all-or-nothing in the Bandung highway network. As in the other papers, goodness of fit was assessed by squared differences from an independently observed OD matrix, and the need for good initial values for coefficients was noted.

Cascetta and Russo (1997) updated an existing set of coefficients, considering them as additional information under maximum likelihood or as Bayesian priors. Normal or Poisson distributions were posited for both the coefficients and the link counts, but randomisations of an artificial five-zone model were specified by the coefficient of variation (standard deviation/mean), suggesting a gamma-like distribution. The four-stage model comprised two purposes, work and school, and three modes – walk, car and bus. Trip distribution was singly constrained with the power of distance as the deterrence function. The utility functions for mode choice comprised time, cost and mode specific constants. Attractions were proportional to a power function of the attractor variables; productions were linear. Cars were assigned with congestion under stochastic user equilibrium, but the spread parameter of the probit distribution was not re-calibrated. Fifteen model parameters were re-calibrated by non-linear generalised least squares. Link flows were related to OD volumes by the assignment vector; the influence of the parameters on the OD volumes was determined numerically by finite differences. Performance was measured by mean square errors of parameters and of link and OD flows, and changes in the objective function. The method was also applied to a small morning peak-hour model of a medium-size town, Reggio Calabria, with initial parameters taken from another town, Parma. The powers of distance in the deterrence functions were around one, except for 0.35 for the journey to work in Reggio Calabria.

Hamerslag and Immers (1988) reviewed the gravity model, as their 'weighted Poisson' model, and matrix estimation by entropy maximisation and information minimisation. The weighted Poisson model could not use aggregate count data, which was considered as either fixed or elastic constraints in matrix estimation. The factors introduced to meet these constraints were considered time and space dependent, making matrix estimation unsuitable for forecasting. These disadvantages were overcome by a 'binary calibration' model, with the structure of a tri-proportional gravity model calibrated by count data under maximum likelihood with Poisson errors. The method had a rather complex structure but had been applied successfully in The Netherlands. The various methods were summarised as shown in table 8.1.

Table 8.1 Summary of four OD estimation techniques

Characteristic	Weighted Poisson	Entropy maximisation information minimisation estimator	Information minimisation elastic model	Binary calibration constraints
Estimation unobserved OD pairs	Yes	No	No	Yes
Apparently contradictory information permitted	No	No	Yes	Yes
Possibility of using traffic counts	No	Yes	Yes	Yes

Characteristic	Weighted Poisson	Entropy maximisation information minimisation estimator	Information minimisation elastic model	Binary calibration constraints
Change of OD pairs	-	Only observed OD pairs	Only observed OD pairs	All OD pairs
Loss of information	No	Yes	Dependent on value of elasticities	No
Time and place dependency of coefficients (if yes, it is not possible to use the model for medium and long term forecasts)	No	Yes	Yes	No
Complex structure of model	No	No	No	Yes

Source: Hamerslag and Immers (1988, table 5)

8.1.3 Matrix estimation from counts

8.1.3.1 Information minimisation and maximum entropy

Van Zuylen and Willumsen (1980) have demonstrated theoretical bases for matrix estimation in information theory and entropy maximisation with similar outcomes. Both were derived from a greatest number of micro-states. In their information minimisation, the state is a vehicle passing a count site; in maximum entropy, it is a vehicle trip from origin to destination, following Wilson's (1969) derivation of trip distribution.

Link counts were taken as fixed constraints, leading to a multi-proportional structure with factors corresponding to counts arising as Lagrangean multipliers in optimisation. OD movements are multiplied by the factors for all the count sites they pass through. In information minimisation, these factors are averaged; in maximum entropy, they are not, giving a bias towards movements counted many times when there has been overall growth from the prior matrix. Van Zuylen (1981) and Bell (1983) recognised this, and added an overall factor to allow for general growth in the matrix.

Inconsistencies between the fixed count constraints would prevent an optimisation converging. Sets of consistent data were estimated under an assumption of Poisson errors in the counts (Van Zuylen and Branston 1982).

8.1.3.2 Generalised least squares – normal errors

Cascetta (1984) allowed for errors in both the prior matrix and the aggregate counts with generalised least squares. This provided maximum likelihood estimates for normal errors. Bell (1984) showed that generalised least squares could approximate to the maximum entropy and information minimisation methods with fixed count constraints by appropriate weighting. Under this approximation, maximum likelihood of Poisson errors in the prior matrix gives the same form as maximum entropy, with a different estimator for the mean whose inverse is the weight – see appendix H. Bell later (1991) provided an algorithm to avoid the negative results that are liable to arise from generalised least squares, particularly for small counts or matrix cell values.

8.1.3.3 Poisson errors

Spiess (1987) found maximum likelihood solutions with Poisson errors in the prior matrix. Counts were initially treated as fixed constraints, but later relaxed to Poisson variables allowing solutions with inconsistent counts. Trip end totals were considered as a special case of constraint, leading to a non-proportional solution which could estimate trips in cells with none in the prior matrix. Under maximum entropy, the solution would be of the multiplicative form:

$T_{ij} = \alpha_i \beta_j t_{ij}$ where t is the observed matrix and T is the estimated matrix

but with a Poisson distribution of t, maximum likelihood is shown to result in the form:

$T_{ij} = t_{ij} / (\rho_{ij} + \alpha_i + \beta_j)$ for $t_{ij} > 0$, where ρ_{ij} is the sampling fraction

and $T_{ij} > 0$ for some $t_{ij} = 0$

The worked example has been fitted by MVESTM, see appendix G.

8.1.4 Other issues in matrix estimation

Matrix estimation has developed in many directions which are not being considered in this study.

8.1.4.1 Congested assignment

Under congested assignment, routings through the network vary with the traffic assigned to it, which in turn affect the estimation of the matrix. This can lead to iteration between assignment and matrix estimation, or the more integrated approaches of Fisk (1988; 1989). Path flow estimation (Bell and Grosso 1998) combines congested assignment and matrix estimation efficiently by processing a limited set of paths.

All work in this study is based on fixed routings.

8.1.4.2 Part route

Matrix estimation has been applied to parts of networks, such as:

- individual junctions
- motorway networks between on- and off-ramps, or
- station-to-station or stop-to-stop movements by train or bus.

Trip distribution is generally concerned with the full journey linking land uses.

8.1.4.3 Dynamic

OD information can be abstracted from short-term changes in traffic flows, eg by matching fluctuations in flow on one approach to a junction with later ones at the exits. Such short-term matrix updating can be used to optimise traffic signal control strategies. Trip distribution generally operates on a much longer timescale.

8.1.5 Software programs

Many transport and traffic modelling software suites include matrix estimation programs, with variations to match the characteristics of the suites and address practical issues in matrix estimation.

OmniTrans works with its Cube structure of demand matrices, recognising dimensions of mode, purpose, vehicle class, period, and sometimes iteration as well as the conventional origins and destinations. Its matrix estimator accepts data aggregated over these dimensions, eg purpose, which cannot be distinguished.

EMME/2 did not include matrix estimation as an intrinsic function. Spiess (1990) developed a macro demadj.mac which utilised its path-skimming functions. It minimises the sum of squared differences between counts and assigned flows, relying on its method of steepest descent to minimise changes in the matrix. The assignment, usually equilibrium, is recalculated at each step. Link or turn counts can be weighted, and parts of the matrix frozen. The procedure has been incorporated in EMME/3. Macros for public transport, accepting link counts on a particular service, and for multiple user classes are also available.

As a traffic model, SATURN works with junction turning counts. With its native equilibrium assignment representing congestion effects, routings change with the matrix, leading to iteration between assignment and matrix estimation. Where parts of the prior matrix are known, they can be 'frozen' so only the other cells are changed to match counts. This allows an incremental approach where just the least reliable cells

are estimated first, with more reliable cells being unfrozen for estimation in later rounds. More recently, it has introduced aggregations of counts (screenlines) and inequality constraints on counts.

Matrix estimation in Contram allows for count observations spanning the short time slices in which the suite works.

VISUM has the TFlowFuzzy procedure for matrix correction, described by Friedrich et al (2000). It can be applied to both highway and public transport models, using count data for individual OD movements, turns, links, initial boardings or final alightings. Intermediate alightings and boardings at transfers between public transport services must be excluded.

Counts are regarded as members of a fuzzy set, within a bandwidth that is symmetrical about the observed value. The set membership function is triangular, with a maximum of 1 at the observed value falling to 0 at the limits of the bandwidth.

The procedure maximises entropy functions that minimise differences between the input and output matrix cells (as per Van Zuylen and Willumsen 1980), and also between the observed counts and the link flows from the new matrix. The latter entropy is calculated from the slack variables, the differences between the flow and the upper and lower limits of the bandwidth.

Bandwidths are set individually as absolute values, but can be calculated as proportions of the observed count. An overall factor can be applied to all bandwidths to ensure all counts fall within them. Parts of the matrix can be frozen, ie excluded from changes, so counts are matched by adjusting the remainder of the matrix. Movements not intercepted at count points can be factored by the average change in counted movements.

VISUM offers two simpler, possibly older, procedures for highway models. 'Path projection' factors all movements passing along one link to match a count. This appears similar to the 'Select link zone factor' function of Netanal. 'Calibrating a matrix' produces factors from a set of counts in the same way and then seeks a set of trip end factors that satisfy the link factors in a Furness-like process. This appears similar to the method used in Tyne & Wear by Irving et al (1986).

HCGMAT is an implementation of the combined calibration method of Gunn et al (1980), which also influenced MVESTM (see section 8.2.1.1). It works at the original sampling level of observations, rather than with grossed up or expanded trips. Gunn et al (1997) described its application to a synthetic prior matrix, using trip end, cost band, link count and roadside interview site matrices to estimate base matrices for the Dutch national model.

8.2 MVESTM

MVA Systematica introduced matrix estimation into its Trips modelling suite a little later than many of its contemporaries, but the program, MVESTM, had a broad range of capabilities, some of which appear to remain unmatched in commercial software.

The Trips suite has since been incorporated into Citlabs' Cube system. MVESTM now also works with the Voyager suite, another component of the Cube system, and is named Cube Analyst. Although this is a fair description of the program's potential that is being explored in this study, the older name is used here for clarity, except for features specific to the Cube Analyst version.

8.2.1 History

MVESTM has a diverse parentage, and occasionally shows it.

8.2.1.1 Combined calibration, RHTM

It drew on the combined calibration method of Gunn et al (1980). This arose in response to mismatches between distribution and generation outputs of the UK Regional Highways Traffic Model (RHTM), derived from roadside and household interviews. It aimed to provide a statistically consistent estimation methodology to calibrate a gravity model, combining information from all pertinent data sets with due regard to their error structure. The gravity model was a partial trip distribution with a tri-proportional deterrence function. The preferred approach was maximum likelihood from Poisson errors, but with scaling between the mean and variance of the distributions to account for errors other than simple sampling, eg in synthesising trip ends. Practical values for such errors and scaling were considered.

Murchland (in Gunn et al 1980) proposed a minimum loss approach focusing on errors in travel cost (trips×cost) rather than trips; the Exponential deterrence function characteristically replicates total travel cost.

8.2.1.2 Tyne & Wear

MVESTM also drew on matrix estimation of the Tyne & Wear (UK) model described by Irving et al (1986). This found a great increase in the number of short-distance trips, passing only one count site, at the expense of longer trips that passed many count sites. Simulations that removed errors of assignment showed a considerable reduction in this phenomenon but did not eliminate it.

The paper suggests there may be insufficient counts to suitably constrain the trip ends of the many zones, particularly in the intermediate and external areas beyond the county being modelled. It seems likely that some movements to or from most zones would still be counted and have factors applied to them, but remaining movements which did not pass through any count site would not be altered. Irving et al (1986) also suggest that the matrix estimation fails to retain the deterrence to travel implicit in the prior matrix.

As an alternative approach, the structure of factors matching the scope of counts was replaced by factors for rows and columns, ie Furnessing, which could be shown to retain the deterrence to travel and allowed factoring of every cell in the matrix. The same structure was used to simulate a target matrix with varied trip end growths from the prior matrix. The target matrix was assigned to give 'observed' link counts for updating the prior matrix.

Row and column factors were derived from the ratios of observed and estimated counts. The ratios were weighted according to the influence of cell values on the Poisson likelihood of the counts. The influence was determined from small perturbations of the cell values. This approach resolved the bias towards shorter trips, but still did not replicate trip ends satisfactorily.

To achieve this, a number of individual cells were 'observed' from the target matrix. In practice, these could be observed by roadside interview. In the matrix estimation procedure, each cell was treated as an additional count. One percent of cells were needed to replicate the target matrix; about 10% were needed for rapid convergence.

Grouping count sites into screenlines was recommended to avoid sensitivity to assignment.

8.2.1.3 Biases

Maher (1987) considered the original algorithm used in Tyne & Wear to be information minimisation. This was not expected to be biased by the number of count points for an overall growth from the prior matrix, unlike maximum entropy methods without Van Zuylen's (1981) overall factor. Maher (1987) confirmed these expectations, but showed that with random variation in the growth rate, information minimisation was biased towards shorter trips as found in Tyne & Wear.

Maher (1987) proposed a two-stage maximum entropy model, first fitting factors for rows and columns (an extension of Van Zuylen's 1981 overall factor) and then factors corresponding to the scope of counts.

This recovered varied trip end growths, unlike maximum entropy algorithms with or without Van Zuylen's (1981) overall factor.

8.2.1.4 Dramatis personae

The development of MVESTM was sponsored by MVA Systematica and The Netherlands Ministry of Transport. The specification was developed by Geoff Copley and Geoff Hyman with assistance from Martin Bach. Miles Logie managed the software development, while George Skrobanski and Al Hynd developed the algorithms and interfacing. Jaap Benschop and Mike Maher sat on the project steering committee. Stuart Tredinnick maintained and updated the programme after its original development. Zhong Zhou has now taken over its development.

8.2.2 Documentation

The theory and practical applications of MVESTM have been published by Logie and Hynd (1990) and Logie (1993).

Electronic documentation is provided with the Cube Analyst software package in two locations:

Trips Help, <Matrix Estimation>

Cube Help, <Cube Analyst>

Although the two are structured differently, much of the material is common to both, including the mathematical background which also appears as an appendix in Logie and Hynd (1990). Only the Trips Help has a bibliography, glossary and index of keywords. The contents of Trips Help appear similar to older printed manuals for MVESTM. References to Help pages, <shown thus>, are to Trips unless stated otherwise.

8.2.2.1 Version

The main analysis of this study used MVESTM version 7, modification 1.8, with library version 7.49. There do not appear to have been any major changes to the core of MVESTM recently; the shell has been updated to integrate it with the Cube suite. Most processing was run with Cube version 4.1.2 or 4.2.3.

8.2.3 Key features

MVESTM takes a statistical approach of maximising likelihoods of Poisson distributions. This is the same basis as GLMs. It can incorporate data from a prior matrix, link counts and trip end totals, allowing for error in any of these sources.

8.2.3.1 Errors in data – confidence and weight

The expected size of error is entered as a 'confidence' in parallel with each item of data. The confidence is the ratio of the mean to the variance, factored by 100.

$$\text{confidence} = 100 \times \text{mean/variance}$$

In the technical documentation of MVESTM, confidence is represented by the symbol λ, but without the factor of 100.

This is the same as the weight applied in GLMs, and is termed w here, so

$$w = \text{mean/variance}$$

MVESTM's 'confidence' is not the same as a statistical confidence interval. In statistical terminology 1/w is an index of dispersion, the variance expressed as a proportion of the mean.

Allowing this proportion to vary from unity turns a Poisson distribution into quasi-Poisson, with greater flexibility to represent various error sources while maintaining the property that its variance is

proportional to the estimated mean. The confidence can be interpreted as an adjustment for expansion of a pure Poisson sampling process. If an item of data has been grossed up from a 7% sample, the confidence is 7; if it is the hourly flow rate averaged from a full two-hour count, the confidence is 200, but only if vehicles do arrive at random, which seems unlikely in practice.

8.2.3.2 Screenlines

The main input to the process that is not treated as uncertain is the routing, which determines which OD movements are observed in a count. In practice, route choice and the resulting proportion of a movement using a link (R_{ijk}) are by no means exact. To minimise errors in routing, the MVESTM methodology is to aggregate counts into screenlines which intercept alternative routes. Ideally, movements between zones on either side of a screenline will cross it once and once only, and no other movement will cross the screenline.

The program will accept a single counted link as a screenline, and recent versions will also handle turns at junctions. However, the nomenclature 'screenline data' is applied to all inputs of aggregated counts. This thesis adopts the same nomenclature for any information introduced in the same way, even when it comprises only a single movement between two zones.

8.2.3.3 Costs

MVESTM can synthesise the prior trip matrix t_{ij} from costs c_{ij} with a Tanner deterrence function

$$t_{ij} = c_{ij}^{-\gamma} \exp(-\lambda c_{ij})$$

This could be calculated to prepare a prior matrix for any matrix estimation program, but in MVESTM the parameters γ and λ can be optimised to give the best fit of the estimated matrix to data. This effectively calibrates the deterrence function, but conventional trip distribution calibration is impeded because only observed trips *or* costs are used by MVESTM.

Observed trip and cost skim matrices can be input together, but information from only one of them is used for any given cell. The trips are used in cells where trips are greater than zero; otherwise the costs are used. This could be used with an observed trip matrix to infill the null cells which could not be observed (eg not intercepted at roadside interview sites), or seed zero cells where there was an (observable) count of zero. However, the cost deterrence function $t_{ij} = c_{ij}^{-\gamma} \exp(-\lambda c_{ij})$ is unscaled by any generation or balancing factor and is unlikely to match the scale of trips in the observed matrix cells, which could be hourly flows or all-day totals between small zones or whole conurbations. Such a scaling difference is not readily resolved within MVESTM, and alternative methods of partial trip distribution and seeding will allow better control.

The standard forms outlined on the Trips help page <MVESTM><Notes on program use><Selection of model form> may also be difficult to implement in practice because of this, and because gravity models generally require balancing factors (parameters a(i) and b(j)) to be free to fit trip ends.

The limitation on inputting either trips or cost as matrices was overcome in this study by presenting the observed trip matrix as a set of screenline count (section 8.3.2) so a fully observed trip matrix could be fitted with a full cost matrix.

There is no allowance for uncertainty in costs.

8.2.3.4 Juxtaposition of structure and information

The matrix is estimated as a multiplicative structure

$$T_{ij} = t_{ij}\, a(i)\, b(j) \prod_k X_k^{R_{ijk}}$$

where $t_{ij} = N_{ij}$ from prior trips

or $= c_{ij}^{-\gamma} \exp(-\lambda c_{ij})$ from costs

Initial trips t_{ij}, either directly entered N_{ij} or calculated from costs c_{ij}, are factored by trip end parameters a(i) and b(j) and by screenline parameters X_k.

The parameters a(i), b(j), X_k, γ and λ are optimised for best fit to the input data. The best fit is defined as minimising an objective function summed from these components:

Component	Objective function	
Cells	$\Sigma_{ij}\, w_{ij}(T_{ij} - N_{ij} \log(w_{ij}T_{ij}))$	if based on trips N_{ij}, not costs c_{ij}
Origins	$\Sigma_i\, w_i(G_i - O_i \log(w_iG_i))$	where $G_i = \Sigma_j T_{ij}$
Destinations	$\Sigma_j\, w_j(A_j - D_j \log(w_jA_j))$	where $A_j = \Sigma_i T_{ij}$
Screenlines	$\Sigma_k\, w_k(V_k - Q_k \log(w_kV_k))$	where $V_k = \Sigma_{ij} T_{ij} R_{ijk}$

There is a general correspondence between items of information and the structure, shown schematically in table 8.2.

Table 8.2 Juxtaposition of data and model structure in MVESTM input files

Information					**Structure**		
Index		Data	Conf	Scope	Index		Parameter
Prior matrix							
Orig	Dest				Orig	Dest	
1	1				1	1	fixed as
:	:				:	:	$t_{ij} = N_{ij}$
i	j	N_{ij}	w_{ij}	*Cell i,j*	i	j	or
:	:	but not c_{ij}			:	:	$t_{ij}= c_{ij}^{-\gamma}\exp(-\lambda c_{ij})$
N_{zone}	N_{zone}				N_{zone}	N_{zone}	
Trip ends				*Implicit*	**Model parameters**		
1					1		a(1) [default=1]
:					:		*Implicit*
origin i		O_i	w_i	*Row i*	i		
:					:		
N_{zone}					N_{zone}		*Row i*
1					N_{zone}+1		
:					:		
destination j		D_j	w_j	*Column j*	N_{zone}+j		b(j)
:					:		:
N_{zone}					$2\times N_{zone}$		b(N_{zone})
Screenlines				**Intercepts**	~		~
Lowest s				Movements ij	$2\times N_{zone}$+1		X_1
:				passing	:		:
screen s		Q_k	w_k	through	$2\times N_{zone}$+k		X_k
:				screen k	:		:
Highest s				Proportions R_{ijk}	$2\times N_{zone}+N_{screen}$		$X_{Nscreen}$
					$2\times N_{zone}+N_{screen}+1$		$-\gamma$ (alpha)
					$2\times N_{zone}+N_{screen}+2$		λ (beta)

Input data and model structure are shown on the left and right respectively. They share the same scopes, which are implicit (shown italicised) for matrices and trip ends, but the movements intercepted by each screenline have to be specified in the intercept file. Double borders outline the contents of input files, which are described in more detail in appendix E.1. The formulations of analyses in section 8.7 and appendix G are set out in this table format.

Although the information input and the form of the model are closely linked by the scopes, they can be specified with great flexibility by using confidences in the data and limits to the parameters.

8.2.3.4.1 Zero confidences in data

As a special case, confidence can be set to zero. Fit to the corresponding data is then ignored in estimating the output matrix.

8.2.3.4.2 Limits to parameters in structure

Initial values and upper and lower limits can be specified for parameters. Parameters can be fixed by setting the upper and lower limits to the initial value.

As a special case, parameters can be fixed at unity (or zero for cost coefficients alpha and beta). This effectively removes the factor from the model's structure.

8.2.3.5 Separation of data and structure

Zero confidences for data and fixed limits for parameters allow the input information and model structure to be separated, as in table 8.3.

Table 8.3 Separation of data and model structure in MVESTM

Screenlines		Intercepts	Model parameters		
Data	Confidence	Scope	Initial Value	Limits	
				Lower	Upper
Q_k	w_k ⇔	Defined by scope of data	*1*	*1*	*1*
			fixed at unity		
Dummy	*0*	Defined by model, eg K factors ⇔	1	small	large
			free to fit model to data		

Italic = null effect on fitted model

The top part of the table shows the introduction of a screenline count Q_k with confidence w_k. The corresponding factor in the model structure is suppressed by setting its parameter to unity initially and constraining it to unity throughout the fitting process.

The bottom of the table shows the definition of a factor in the model structure whose parameter can vary to find the best fit to data. The corresponding data is ignored because its confidence is set to zero.

Data (with a confidence) can be introduced as any linear combination of matrix cells by specifying the intercept proportions R_{ijk} as coefficients. The scope of parameters in the model structure can be defined similarly. This gives great flexibility in fitting models to data.

8.2.3.6 Statistical measures

The MVESTM suite offers the main statistical outputs that may be expected from model fitting:

- fit to observations, as listings against estimates with differences in the print file
- overall fit of the model, as the final objective function in the execution log file
- standard errors of model parameters
- accuracy of predictions, as the sensitivity matrix.

The formulations correspond to measures available from a GLM, with their attendant approximations to known distributions, but these statistical interpretations are not always clearly documented. Standard errors and the sensitivity matrix are produced by the auxiliary program MVESTE, but these outputs are not all as expected – see appendix F.

8.3 Application of MVESTM

8.3.1 Trip distribution model structure

The model structure of MVESTM is

$$T_{ij} = t_{ij}\, a(i)\, b(j)\, \Pi_k X_k^{R_{ijk}}$$

where $t_{ij} = N_{ij}$ from prior trips

or $= c_{ij}^{-\gamma} \exp(-\lambda c_{ij})$ from costs

A trip distribution model is specified by fixing the screenline parameters X_k at unity and presenting a cost matrix c_{ij} to provide the initial trip matrix t_{ij}. The structure then takes the form of a trip distribution model

$$T_{ij} = a(i)\, b(j)\, c_{ij}^{-\gamma} \exp(-\lambda c_{ij})$$

The trip end parameters a(i) and b(j) are left free to fit the products of balancing factors and trip end totals, whether or not trip end data is entered. Optimising both cost parameters γ and λ calibrates a Tanner deterrence function. Fixing γ or λ at zero (not the MVESTM initial value of unity) calibrates an Exponential or Power function respectively. Fixing both γ and λ at zero suppresses all cost deterrence effects, providing a null model from which to gauge the importance of trip distribution.

8.3.2 Entry of trip information as screenline data.

Conventional calibration of a trip distribution takes both a cost matrix and a trip matrix, and compares costs and trips in corresponding cells. For any one cell, MVESTM takes information from only one matrix. Since costs are entered as a matrix, trip information has to be presented as screenline data.

For a trip matrix with information disaggregate to the zonal level, the trips in each cell are presented as the count at one screenline. The screenline has to be described in the intercept file as intercepting the one zone-to-zone movement represented by that cell.

In the initial small-scale trial of MVESTM for conventional calibration, a desire-line network was built with direct links between every pair of zones. These links all had the same cost, so each carried only the one direct zone-to-zone movement under all-or-nothing assignment. Each link was specified as a separate screenline, with a count equal to the trips for that movement. Intercept and screenline files could then be generated by the standard procedures for preparing MVESTM data. The method was convoluted even on a small scale, and complicated by practical problems such as intrazonal movements.

8.3.3 Intercept file structure

The owners of MVESTM, Citilabs, supplied a document 'InterceptFile.doc', which describes the binary data structure of the intercept file. This allowed the scope of screenlines to be specified directly, rather than by processing paths through a network.

More detail is given in appendix E, section E.1.6.

8.3.4 Source code for the MVESTM shell

Citilabs granted a licence for their software that included Fortran source code for the shell of the MVESTM; this allowed its inputs and outputs to be modified. The core optimiser was supplied as a compiled object, which could not be examined or modified, to protect Citilabs proprietary interests.

8.3.5 Deterrence functions

The three main analytic deterrence functions were considered – Exponential, Tanner and Power. The source code was amended to allow a negative coefficient of alpha, the cost coefficient for the Power function. This gives a sensible form for the Power function, with travel diminishing as cost increases, and the concave form of the Tanner function which had been found to give a better fit to the Wellington data.

8.3.6 Initial values and convergence

MVESTM accepts initial values of parameters as starting points in its search for the optimal values; by default these are unity. Preliminary runs showed that convergence could be difficult from these defaults, or null initial values of zero for the cost parameters. They were therefore given initial values in the correct range, based on earlier calibration by GLM.

Table 8.4 Initial cost parameter values

Deterrence function	Coefficients calibrated by GLM		Initial parameter values set for MVESTM	
	Log(cost)	Cost	γ	λ
Exponential	~	0.0637	~	0.05
Tanner	0.645	0.0364	0.5	0.03
Power	1.398	~	1.5	~

Estimated matrices generally corresponded closely with trip distributions fitted or synthesised by other methods, typically within 0.01 trips. However, there were a few larger differences, perhaps of 10 or so trips, in a few cells of the matrix. These appeared to cluster in the remote corners of the study area, in the Wairarapa and Kapiti Coast. Convergence difficulties were also noted in these areas when synthesising trip distribution with Trips' MVGRAM and perturbation testing with Voyager's Fratar programs. On a practical scale, such exceptions are small. They would be unnoticeable in most real matrices; they only stand out against pure synthesised trip distributions.

It was found that these convergence problems could generally be resolved by setting good initial values for the trip end parameters. These were taken from the exact values fitted in calibration on the full HIS matrix by GLM; there was thus a different set for each deterrence function. Scalings were set to the same value, though this did not seem to resolve the convergence problem.

These initial values are probably better than are available in practice, but they have been adopted to focus on the applications of matrix estimation techniques rather than computational issues. The exact

parameters from the full HIS trip matrix also appeared to give adequate convergence when calibrating on actual screenline counts, for which the final parameters would differ.

Convergence issues were also manifested in the number of iterations and by the optimiser fixing parameters because they were not contributing to the estimation. These parameters often appeared to be for zones:

- with small numbers of observed trip ends
- in remote parts of the study area, as above
- in exclaves where screenlines differ from natural boundaries (section 8.7.7.7.1).

Convergence generally appeared better with:

- the Exponential deterrence function rather than the Power
- confidences around MVESTM's default of 100 rather than the 0.6 applied to the HIS
- ordinary trip end confidences, rather than weak.

Convergence problems can occur in trip distribution models where cost effects are very significant by usual standards, with changes in deviance up to 10 but small by comparison with many models fitted here, with changes in hundreds.

MVESTM offers several options for controlling the optimisation process. Setting more frequent recalculations of the Hessian with the ITERH parameter often reduced the number of iterations, but otherwise no clear benefit was found in altering the default settings. The defaults were taken as standard.

8.3.7 Empty zones and zero counts

The rows and columns of the estimated matrix for empty zones, without any trip ends in the input, were set exactly to zero by setting costs and their confidences in the prior matrix to zero. The corresponding trip end parameters were fixed at unity.

Seeding the zero trip end counts with small numbers appeared to be a practical alternative, but left minor perturbations (and problems with MVESTE if the corresponding parameters were not fixed). The trip end parameters could not be fixed at zero because their logarithms are taken before entry into the optimiser.

MVESTM's original algorithm to aggregate link counts within a screenline set the screenline's confidence to zero if all the link counts were zero. This is sensible in practice where it is unlikely to have links without traffic in a network, but many cells in a sparse matrix have a valid sample of zero, observed with the same confidence as other cells. The algorithm was changed to retain the confidence input in the screenline file.

8.3.8 Computing environment

The routines to prepare data, run MVESTM and process its output were developed and run in the graphical interface of Cube's Application Manager, previously known as TRIPSWIN. Alternative formulations were defined, run and stored with Cube's Scenario Manager. Initially the subsidiary processing was by Trips routines, from MVESTM's original software package, but later Voyager routines and scripts were also used, taking advantage of the Voyager matrix format's double precision.

The Cube operating environment with the Trips and Voyager routines were supplied gratis by Citilabs.

8.4 Statistical interpretation

MVESTM is founded on a sound statistical model and its outputs can be interpreted in the light of statistical theory which it shares with GLMs. It can provide useful measures of model fit and significance but to do so it needs a proper scaling of errors. This scaling might be derived either from the internal consistency of components, reflected in residuals, or by setting the weighting (in MVESTM terms 'confidences') of inputs to the true scale of their variability as determined externally.

8.4.1 Deviance and the objective function

The main statistical measure used is the deviance, a function of the likelihood. The same likelihood is maximised by MVESTM, but its objective function (`FBEST` in the execution log) omits terms which are fixed for a given dataset. The deviance is formulated so that it is zero for an exact fit to observations and is the same as the sum of squared residuals in simple regression, with similar properties.

$$\text{Deviance} = -2(\ \text{FBEST} - \sum(wH - wH\log(wH))\ )$$

where, in MVESTM terminology

$$\text{objective function FBEST} = \sum(\ wh - (wH)\log(wh)\ )$$

H is observed data

h is estimated data, and

w is its weight, or confidence/100

8.4.2 Change in deviance

The significance of trip distribution is measured by the reduction in deviance, hence improvement in fit, when cost components are added in to the model. This requires the fitting of a null initial model, without cost components, as well as a trip distribution model including them.

Without cost components, the initial model is a 'flat' matrix, with cells simply proportioned by its trip ends. These trip ends do not necessarily match the input trip ends, because the null model may also be fitting around screenline counts or segment totals which are not consistent with a flat matrix built on the input trip ends.

This inconsistency is reflected in the deviance of the null model. In many cases, noted below, the trip distribution model is determinate with no deviance, or with relatively small deviance. The change in deviance which measures the significance of trip distribution then arises mainly, or wholly, from the inconsistency of the data with the initial flat model. This is considered in more detail in section 8.5.4.

Since the change of deviance is proportional to sample size, *ceteris paribus*, it is interpreted as an amount of information about trip distribution that can be abstracted from various datasets and formulations.

8.4.3 Residual deviance

Changes in deviance are not scaled by residual deviances. This gives a χ^2 statistic rather than an F ratio. It follows the practice adopted in this study because the expected residual differs from unity due to sparseness in the data. This sparsity in matrix cells no longer occurs with aggregation to large segments, counts of traffic across screenlines or trip ends. However, there is still no reliable residual deviance in many cases:

- Calibration on a single matrix segment or individual or aggregated screenline count gives a final trip distribution model that is just determinate. No residual deviance is expected from a fully converged analysis.
- Coarse aggregation to large segments leaves few degrees of freedom. Aggregation to 3 x 3 segments leaves only two degrees of freedom after fitting the two parameters for a Tanner model. Practical cases of a single screenline forming 2 x 2 quadrants leave a single degree of freedom only if the two intersector quadrants can be counted separately.
- Even when counts at four screenlines in three periods and two directions are entered separately leaving up to 23 degrees of freedom, some of the residual deviance may arise in strata that reflect more on the consistency of period and direction factoring than of trip distribution.

Residuals that are available will be different for every model. Leaving changes in deviance unscaled allows the changes to be compared on a common scale.

8.4.4 Confidences/weighting

In the absence of scaling by residuals, the scale of deviance is determined by the confidences or weighting of the input data.

MVESTM is based on Poisson likelihoods, with variances proportional to means; the proportion is entered as 'confidence' (100×mean/variance). This has the convenient property that the confidence remains the same for aggregations of independent counts with equal confidences, since the variance of a sum of independent random variables is the sum of their variances.

8.4.4.1 Observed HIS trip matrix and trip ends

The main source of error in this dataset is taken to be random sampling, which does produce a Poisson-like error. The confidence is based on the overall expansion factor of 157.9 from the sample of home-work pairings to a trip matrix representing the whole population. The confidence is thus 100/157.9 or 0.63331, equivalent to the weighting for GLMs.

This is applied to cells and segments of the trip matrix and as the 'ordinary' confidence for trip ends derived from the matrix.

8.4.4.2 Screenline counts

Uncertainty in the fit of screenline counts does not arise just from the same sampling process. If there were just a simple random Poisson process in counts of vehicles, long-term traffic surveys could be large enough to reduce the consequent error to a negligible level. A count of 10,000 vehicles – just one full day on one major link – would leave a sampling error of only 1%.

Day-to-day and seasonal variations present more complex patterns, some parts being systematic. Residual random effects in the underlying mean flows are not readily quantified, and may still only be a minor part of the differences between observations and model estimates. Specification and other modelling errors may be the major components.

Expectations for the whole of the differences between traffic counts and model estimates have been derived from assignment validation criteria. These have been taken from the *Economic evaluation manual* (EEM) (NZTA 2008) and the UK *Design manual for roads and bridges* (DMRB) (Highways Agency 1996). These offer several measures of fit; attention has focused on those based on the GEH

$$\sqrt{((\text{model}-\text{observation})^2/(\text{model}+\text{observation})/2)}$$

An approximate relationship between the GEH criteria and the MVESTM confidence is derived in appendix I and the results are shown in table 8.5.

Worksheet 8.4 of the EEM gives two exact criteria for individual links:

- At least 60% of individual link flows should have GEH less than 5.0.
- At least 95% of individual link flows should have GEH less than 10.0.

EEM gives two further criteria which need interpretation for probabilistic calculation:

- All individual link flows should have GEH less than 12.0 (EEM).
  For table 8.5, this is based on about 100 counted one-way links in the WTSM, with three periods but only one class of traffic count, needing a 99.9% level to be sure of all 300 link flows.
- Screenline flows should have GEH less than 4.0 in most cases (EEM).
  Probabilities of 80% and 95% are tabulated to represent this criterion.

Table 8.5 Confidences from GEH validation criteria

Source	Probability P	Standard deviates $\Phi^{-1}((1+P)/2)$	GEH	MVESTM confidence
EEM links	60%	0.842	5.0	2.83
EEM links	95%	1.960	10.0	3.84
EEM links	All *99.9%*	3.290	12.0	7.52
EEM screenlines	Most *80%*	1.281	4.0	10.26
EEM screenlines	Most *95%*	1.960	4.0	24.01
DMRB 5i links	85%	1.439	5	8.29
DMRB 5ii screenlines	All *99%*	2.576	4	41.47
DMRB 5ii screenlines	(or almost all) *90%*	1.645	4	16.91

Percentiles shown in italics are interpretations of the textual criteria alongside them, used for calculation

The DMRB, volume 12.2.1, table 4.2 also sets out criteria for validation in terms of GEH:

- At least 85% of individual flows are expected to have a GEH of less than 5.
- 'All (or almost all)' screenline totals are expected to have a GEH of less than 4.

There will be fewer screenlines than individual links, particularly under the DMRB's expectation that screenlines normally comprise more than five links. Of the 17 two-way screenlines defined for the WTSM, only one meets this expectation; another two have exactly five links, but four are only a single link. Only one of the four screenlines defined for this study has more than five links. This may reflect the geographic constraint of the Wellington region's network. Under these circumstances, the DMRB may require a similar probability to the EEM's for 'most cases', taken here as 80%–95%.

The lower GEH expected of screenlines compared with individual links argues against the independence of links within a screenline, and for a negative correlation. This is consistent with routing error where flow lost on one link is likely to be gained on parallel links.

The variances are between 2.4 and 35 times those expected from counting traffic for one hour if vehicle arrivals followed a random Poisson process, which would give a confidence of 100. This argues against the main source of error arising from such sampling, or for it following a Poisson form for that reason.

Both the EEM and the DMRB present other criteria based on percentage errors. This suggests that the standard error, rather than the variance, is proportional to the mean. The coefficient of variation can be converted to a confidence at a given volume, but this produces a different confidence for each count.

EEM also gives criteria for the coefficient of determination (R^2) of regression on a scatterplot. This suggests that the error is a constant across links and its size is related to the range of flows plotted. However, there is no indication that this measure, percentage errors or the GEH reflect any clear expectation or knowledge of the way errors vary with flows, or that they are any more than convenient ways of expressing errors.

In both the EEM and the DMRB, GEHs are specified for one-way, hourly flows typical of traffic assignment models. The DMRB criteria are for validating traffic assignment models rather than travel demand models. Matrix validation is addressed separately in section 4.3.42 of the DMRB, but no numerical criteria are specified there, or in the strategic demand model checks in worksheet 8.5 of the EEM.

The dispersion of differences between the hourly screenline counts and the WTSM estimates was modelled briefly by HGLM. For the HBW car trip crossings of the four screenlines selected for this study, the best fitting form was

$$\text{Variance} \propto \text{mean}^{1.2}$$

This is closer to the Poisson/GEH form with a power of 1 than the gamma form with a power of 2 that gives a constant coefficient of variation or percentage error. The fitted form was not a significant improvement on the Poisson form, which gives an initial confidence 5 or 6. There appear to be systematic differences between screenlines (across periods and directions); the confidence improves to 20 by fitting them as fixed or random effects.

Based on this and consideration of table 8.5, a common confidence of 6.0671 has been adopted for all screenline counts and their aggregates. (It was originally a round value of 5 in an earlier weighting scheme.) This is for ease of computation and interpretation, rather than from a knowledge of the form and interdependence of the errors. Such information goes to the core of transportation modelling and is too large and complex a topic to address in this study.

Counts are presented as hourly flows as per the EEM, avoiding assumptions of independence of hourly counts within one period. With an assumption of independence between the three periods, the aggregate over them is

$$1 \text{ hour of AM flow} + 1 \text{ hour of IP flow} + 1 \text{ hour of PM flow}$$

Aggregations over screenlines and directions are similarly summations of hourly flows.

8.4.4.3 Synthetic trip ends

To complement screenline counts without recourse to a fully observed matrix, synthetic trip ends were taken from the WTSM trip generation models.

There are no standards for validating synthesised zonal trip ends in the EEM or the DMRB. A reference to RHTM by Gunn et al (1980) suggests a coefficient of variation of 0.3. For a typical WTSM zone with 1000 24-hour HBW car trips, this gives a standard error of $0.3 \times 1000 = 300$, or a variance of $300 \times 300 = 90{,}000$. This is 90 times the mean of 1000, giving an MVESTM confidence of $100/90 \approx 1$. Based on an RHTM zone say 10 times bigger, the confidence will be 10 times smaller, ~0.1.

In the WTSM, productions were modelled at the person level, and attractions were modelled between 74 sectors of up to five zones, with correction factors for TLAs. The reported measures of fit were difficult to interpret as residual model error at the zonal level.

The WTSM base year synthesised 24-hour HBW car trip ends were compared with those observed in the HIS. For productions, this suggested a residual modelling error similar in scale to the sampling error of the survey. The form of the error appeared more like the gamma, with a constant coefficient of variation, than the Poisson, with a constant confidence. For attractions, there appeared to be little error beyond that of sampling, contrary to the accepted wisdom that productions are modelled better than attractions.

In the absence of better information, the confidence for all synthetic trip ends was kept at that used for trip ends observed in the household survey, 0.63331. This falls into the range suggested by the coefficient of variation from RHTM and simplifies comparison of calibrations on different datasets. It is treated as a Poisson process for simplicity in coding and interpreting MVESTM.

8.4.4.4 Strong, ordinary and weak trip end confidences

As a test of sensitivity to trip end confidences, analyses were run with 'strong' and 'weak' trip end confidences as well as the 'ordinary' confidence, set above at 0.63331. Strong confidences are 100 times larger than ordinary confidences; they approximate to trip ends acting as fixed constraints. Weak confidences are 100 times smaller than ordinary; trip end data become a secondary source of information for allocating trips between zones where there is no evidence from screenline counts.

8.4.5 MVESTM practice

MVESTM documentation does not advise on absolute levels for confidences, but suggests setting their relative levels in accordance with the reliability of different datasets or sources. No interpretation is given of the objective function (`FBEST`) as a likelihood which can be tested statistically.

No setting of 'real' confidences to allow statistical interpretation is known in practice. Confidences may be set with order of magnitude differences reflecting relative importances, or by an empirical search for values that give the best results, often in terms of the validation criteria discussed above.

8.4.6 Other measures

The standard errors of parameters and accuracy of predictions (matrix sensitivity) output by the subsidiary programme MVESTE could not be reconciled with expectations for very simple models (see appendix F). Consideration was given to deriving parameter variances from the Hessian matrix returned by the optimiser, or by perturbing the estimated parameter values. However, the ordering of the Hessian was unclear and perturbation required extra matrix estimation runs.

All these measures are scaled to input confidences as MVESTM does not adjust to residuals.

8.5 Trip distribution information available from screenline counts

This section considers the ability to abstract trip distribution information from screenline counts and in particular the role of trip end data.

8.5.1 Single screenline

Consider the simplest case of a study area divided by one screenline into two sectors. The PA matrix is thus divided into four segments, or quadrants. In the first instance, this matrix is also taken to be an OD matrix with all journeys from home to work, which is quite typical of the morning peak.

Total trips for each segment and the sector trip ends are denoted as follows:

Prod\Attr	1	2	Trip ends
1	T_{11}	T_{12}	T_{1*}
2	T_{21}	T_{22}	T_{2*}
Trip ends	T_{*1}	T_{*2}	T_{**}

For each segment there is a cost C. No general aggregate of zone-to-zone costs within a segment has been found to represent all their distribution effects, or seems likely to exist. Exact cases can be made:

- if each zone-to-zone cost in a segment is the same, or
- for a 2x2 zone subset of the full matrix, taking one zone to represent each sector. (This assumes that individual zone-to-zone movements are distinguished at the screenline, as in a roadside interview survey rather than a simple traffic count.)

The formula

$$-\lambda \times (C_{11} - C_{12} - C_{21} + C_{22}) = \log T_{11} - \log T_{12} - \log T_{21} + \log T_{22}$$

$$= \log(T_{11} \times T_{22} / T_{12} \times T_{21})$$

developed in section 3.1 applies in these special cases, and is taken as indicative of the relationships in more general cases of aggregation.

If trips T and costs C are known for each of the four segments, there is just sufficient information to calibrate the cost coefficient λ.

Prod\Attr	1	2
1	T_{11}	T_{12}
2	T_{21}	T_{22}

Trailing diagonal intersector

Leading diagonal intrasector

λ cannot be determined unless there is a difference between the costs on the leading diagonal, $C_{11} + C_{22}$, and the costs on the trailing diagonal, $C_{12} + C_{21}$. This will usually be the case because the leading diagonal represents intrasector movements, which will generally be shorter than the intersector movements of the trailing diagonal:

$$C_{11} + C_{22} < C_{12} + C_{21}$$

8.5.1.1 Single quadrant

Only the intersector movements cross the screenline, so counts there can only observe trips for the two quadrants on the trailing diagonal, T_{12} and T_{21}. The screenline counts alone are insufficient to calibrate the trip distribution.

However, if trip ends are known for each sector, the trips for each quadrant can then be derived from a screenline count of trips in just one quadrant. If, say, T_{21} is observed as the screenline crossings from sector 2 to sector 1 in the morning peak, then

$$T_{11} = T_{*1} - T_{21}$$

$$T_{22} = T_{2*} - T_{21}$$

$$T_{12} = T_{**} - T_{*1} - T_{2*} + T_{21}$$

This is just sufficient to calibrate a trip distribution model, but it needs a trip generation model or other source of trip end totals.

8.5.1.2 Two separate quadrants

Movements in two of the quadrants can be observed at screenlines. Totals for each might be observed separately, as counts by direction of crossing. If trip ends are also available, there is now one degree of redundancy. This gives indeterminate results until a further criterion such as maximum likelihood is applied. It also provides an internal measure of fit of the data to the model.

Alternatively, it might be used to calibrate a generation model if one is not available to provide sector trip ends. These can be modelled from production planning data P, eg households, and attraction data A, eg workplaces or workspace. Trip rates p and a have to be calibrated to give models of the form

$$T_{i*} = p \times P_i$$

$$T_{*j} = a \times A_j$$

The two trip rates are related through the global totals of trips and planning data

$$T_{**} = p \times P_* = a \times A_*$$

so only one degree of freedom is needed to find both. Substituting this relationship into alternative formulations for the trips in an unobserved segment

$$T_{11} = p \times P_1 - T_{12} = a \times A_1 - T_{21}$$

gives formulae for the trip rates in terms of the planning data and observable trips

$$p \times (P_1/P_* - A_1/A_*) = (T_{12} - T_{21}) / P_*$$

$$a \times (A_1/A_* - P_1/P_*) = (T_{21} - T_{12}) / A_*$$

The formulae relate the imbalance in generators (production and attraction) in the sectors to the net movement in trips between them; the costs and trip distribution are not involved. If the productions and attractions balanced exactly in each sector, there could be any number of intrasector trips, rendering the trip rates and trip distribution indeterminate.

8.5.1.3 Two quadrants aggregated

While there are strong patterns of travelling from home in the morning and returning in the evening, the totals for the two observable quadrants of a PA matrix cannot be completely distinguished by the direction, or any other observable property, in simple traffic counts. However, as a worst case, the sum of counts in both directions will equal the sum of the two PA quadrants over a whole day.

This loses one degree of freedom, so a single trip distribution parameter is again just determinate if trip ends are known. Planning data is not sufficient because the net movement between sectors, $T_{12}-T_{21}$, is no longer available to determine trip rates.

From the combined screenline observation

$$T_S = T_{12} + T_{21}$$

and sector trip ends, individual quadrant trips can be derived as:

$$T_{ij} = T_{i*} + T_{*i} - T_S \qquad \text{for intrasector, } i = j$$

$$\text{or } T_{i*} - T_{*i} + T_S \qquad \text{for intersector, } i \neq j$$

8.5.2 Multiple screenlines

At a single screenline, there is insufficient information to calibrate a trip distribution from counts of screenline crossings alone; sector trip end totals are also needed.

Counts at multiple screenlines provide more information about different segments of the PA matrix. However, the screenlines also break the study area up into more sectors.

8.5.2.1 Sector topology

In general, N screenlines will divide a study area into at least N+1 distinct sectors. There could be more if screenlines cross or screenlines form cordons which intercept through movements.

To fit interaction effects such as trip distribution, main effects must be fitted. This will require one production trip end total and one attraction trip end total for each sector, less one common global total, needing $2(N + 1)-1 = 2N + 1$ items of information. With at least one item of information to determine the interaction, $2N + 2$ items are needed.

This is not available from just one item of information per screenline, such as a two-way, all-day traffic count. Even if separate period and direction counts yield two distinct items of information, ie separate trip totals for the two observable quadrants, this is still insufficient to calibrate trip distribution without trip end information.

In the practical example of Wellington (figure 8.4) there are at least five distinct sectors between the four screenlines, possibly seven if south, west and north Wellington are distinguished by their routing through the central cordon. This will always be too many sectors to establish production and attraction trip ends if just one or two items of information are available from counting at each screenline, but within the 24 degrees of freedom available if (three) period and (two) direction counts are truly distinct in the information they provide about the PA matrix.

8.5.2.2 Segment determinacy

Ignoring for the moment the topology of the study area, N counts might uniquely determine the trip totals for N segments. These must define at least $\sqrt{N}$ sectors, requiring at least $2\sqrt{N}$ items of information to determine the main effects plus one interaction effect. Since $N > = 2\sqrt{N}$ for $N > = 4$, this is possible for any non-trivial case with two or more sectors and the margin of determinacy increases with N. At $N = 6$, the margin allows for the three intrasector segments in a 3 x 3 sector matrix to be unobservable.

The simplest relationship between counts and segment totals is a direct one, ie each segment is somehow counted directly. This reduces the analysis to that of a fully observed matrix at the sector level, equivalent to conventional calibration of a zonal matrix, subject to the aggregation of costs over a segment.

This requires that each count is specific not only to a single origin sector, but also to a specific destination sector. This is readily plausible if and only if every sector touches every other sector and traffic can be counted at their boundaries. This limits the number of sectors to four under simple two-dimensional topologies. Otherwise movements to nearer and further sectors are liable to be aggregated into the same count. This might work for a point-to-point airline network without hubbing, but grade-separated road networks rarely help distinguish movements.

N segment totals could also be uniquely determined by N counts through a non-degenerate set of linear equations. This allows counts to comprise several sector-to-sector movements. However, the possible sets of equations will be limited by the practical expectation that any sector-to-sector movement will be intercepted only once by a screenline, or at most twice for through movements in and out of a cordon. The data matrix must therefore comprise 0, 1, or possibly 2, akin to an ‘indicator’ matrix for dummy variables.

This presents an interesting theoretical exercise in topology and linear equations to find configurations of screenlines at which crossing counts alone can determine the trip distribution. Robillard (1975, lemma 3) offers a necessary and sufficient condition to be able to estimate trip ends from link counts, with the simple corollary that the number of link counts must at least equal the number of trip ends, less one for the equality in production and attraction totals.

However, the preceding consideration of sector topology suggests that trip end totals will also be needed in practical cases. It has been shown that these can be derived from planning data at the expense of one item of screenline count information.

8.5.3 Within-sector trip end patterns

Up to this point, the calibration of trip distribution has depended only on the total trip ends for each sector and not on the pattern of trip ends within a sector. However, this interpretation is based upon an aggregate cost C for each segment to represent all trip distribution effects.

Section 7.6 suggests that no such single aggregate measure exists. If trip distribution effects cannot be completely related to a single summary cost for each segment, then they still depend to some extent on the pattern of individual zone-to-zone costs within the segment. This pattern of costs within the segment is weighted by the pattern of trip ends within its defining sectors. The patterns of trip ends within sectors will then affect the calibration of trip distribution.

Empirical findings suggest so.

8.5.3.1 Hierarchy in aggregation

At first sight, calibrating trip distribution by matrix estimation on aggregate data appears to complement methods developed in spatial analysis exactly.

Matrix estimation fits a trip distribution to between-segment contrasts, without any arbitrary aggregation of costs – these are still presented and processed as individual zone-to-zone values. This is the upper level of a (spatially) hierarchical trip distribution. Information about the lower level, the contrasts within segments, is lost in aggregation of trips by segment. This appears to complement spatial analyses that fit trip distributions to the lower level, within segments, while the upper level information is absorbed by K factors for the segments.

Further effects may arise from an intermediate level in the hierarchy, which is trip distribution information arising from contrasts between four-square sets of zones within the same production sector but split between attraction sectors, or vice versa. Figure 8.1 shows such a set of zones within a matrix.

Figure 8.1 Four-square set of zones in an intermediate level of hierarchy

Sector	~	Sector B		~
Zone		Zone 1	Zone 2	
:				
Sector A		Lower cost	Higher cost	
:				
Sector C		Higher cost	Lower cost	
:				

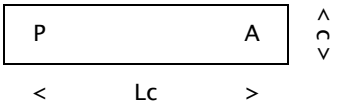
This pattern of costs could occur with sectors A, B and C lying along a corridor; within sector B, zone 1 is closer to sector A and zone 2 is closer to sector C.

In a distribution model, trips to zone 1 will tend to come from sector A, and trips to zone 2 will tend to come from sector C because of the lower costs. The allocation of attraction trip ends between zones 1 and 2 within sector B can thus affect the allocation of trips between segments A→B and C→B. The scale of this effect depends on the cost coefficient. Hence there is a linkage between zonal trip ends within a sector and the cost coefficient while fitting to aggregate segment totals.

There would be no linkage if all the zone-to-zone costs within a segment were the same, allowing a single value of cost to represent the whole segment exactly.

8.5.3.2 Separation of productions and attractions within a sector

This is a set of minimal cases of effects in the intermediate hierarchy. Consider long, thin sectors with productions P and attractions A at opposite ends. The width, c in generalised cost, determines the minimum costs to adjacent sectors. Internally, the productions and attractions are separated by Lc, where L>1 for this discussion.



Two such sectors with productions and attractions all equal are adjacent and separated by a screenline. The proportion of trips crossing the screenline is p. This can be expressed as

$$p = (T_{12} + T_{21})/T_{**}$$

$$T_{12} = T_{21} = p\,T_{**}/2 \qquad \text{by symmetry}$$

$$T_{11} = T_{22} = (1-p)T_{**}/2 \qquad \text{by subtraction}$$

in the terms of section 3.1:

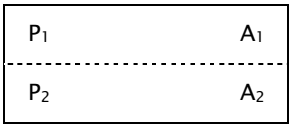
$$-\lambda \times (C_{11} - C_{12} - C_{21} + C_{22}) = \log(T_{11}\times T_{22} \,/\, T_{12}\times T_{21})$$

$$= 2\times\log(\,(1-p)\;/\;p\;)$$

$$p = 1/(1+\exp(-\lambda \times (C_{11} - C_{12} - C_{21} + C_{22})\,/\,2))$$

Case A

With sectors side-by-side and productions or attractions at adjacent ends, the costs for all PA pairings are approximately the same, Lc. With trips distributed equally to each pairing, half will cross the screenline.



Taking all costs to be the same, the cost matrix is:

Lc	Lc
Lc	Lc

$$P = 1/(1+\exp(-\lambda \times (C_{11} - C_{12} - C_{21} + C_{22})\,/\,2))$$

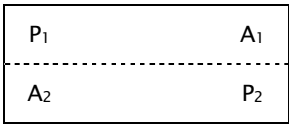
$$= 1/(1+\exp(-\lambda \times (Lc - Lc - Lc + Lc)/\,2))$$

$$= 1/(1+\exp(-\lambda \times 0))$$

$$= 1/2, \quad \text{irrespective of the cost coefficient } \lambda.$$

Case B

With production and attractions reversed in one of the sectors, the costs of the between-sector PA pairings ($P_1 \rightarrow A_2$, $P_2 \rightarrow A_1$) are less than the costs within sector ($P_1 \rightarrow A_1$, $P_2 \rightarrow A_2$), so more than half the trips will cross the screenline.



The cost matrix is unusual in that the intrasector costs on the leading diagonal are larger than the intersector costs on the trailing diagonal.

Lc	c
c	Lc

$$p = 1/(1 + \exp(-\lambda \times (Lc - c - c + Lc)/\ 2))$$

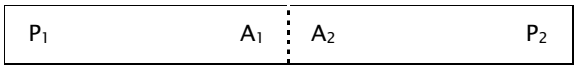
$$= 1/(1 + \exp(-\lambda c \times (L-1)))$$

$$> 1/2 \quad \text{for +ve } \lambda \text{ and } L>1$$

Screenline crossings are related to both the trip distribution, represented by the cost coefficient λ, and the internal distribution of generations, represented by L. This is still the case for L<1; the cost matrix then takes the more usual form with lower costs within the sectors and p<1/2.

Case C

With sectors end to end and both productions or both attractions in the adjacent ends, internal costs within sectors will be lower than between-sector costs so less than half the trips will cross the screenline.



The cost matrix is:

Lc	Lc+c
Lc+c	Lc

$$p = 1/(1 + \exp(-\lambda \times (Lc - (Lc + c) - (Lc + c) + Lc)/\ 2))$$

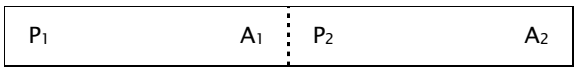
$$= 1/(1 + \exp(\lambda c))$$

$$< 1/2 \text{ for + ve } \lambda$$

The screenline crossings depend on the minimum separation between sectors but not on the relative separation of productions and attractions within sectors.

Case D

With a production and an attraction in the adjacent ends, the cost between them is less than the internal costs. However, the other between-sector cost between the far ends is higher, so the outcome is not obvious.



The cost matrix is:

Lc	2Lc+c
c	Lc

$$p = 1/(1 + \exp(-\lambda \times (Lc - (2Lc+c) - c + Lc)/ 2))$$

$$= 1/(1 + \exp(\lambda c))$$

This is the same as for the previous configuration. If, say, there is a direct link or service between the far ends the cost is reduced by Dc. The cost matrix is then:

Lc	2Lc+c-Dc
c	Lc

$$p = 1/(1 + \exp(-\lambda \times (Lc - (2Lc+c-Dc) - c + Lc)/ 2))$$

$$= 1/(1 + \exp(-\lambda \times (-2c+Dc)/ 2))$$

$$= 1/(1 + \exp(\lambda c(1-D/2)))$$

D reduces the effect of the between-sector separation until at D=2, half the trips cross the screenline. At D = 2L, the configuration reverts to that of case (B) above with the far ends folded together.

Case (B) shows that the internal distribution of generators within a sector can affect screenline crossings and the effect is related to the cost coefficient λ. However, the other cases show that the effects of internal separations tend to cancel.

8.5.4 Amount of information about trip distribution

The statistical significance of trip distribution effects can be measured by the change in deviance when the cost component is added to the model. The change in deviance is proportional to the sample size in a simple experiment, so it can be thought of as a measure of information that can be abstracted from different subsets or aggregates of data.

If the final model including trip distribution effects is only just determinate as in several cases considered above, the residual deviance will be zero. The reduction in deviance is then equal to the initial deviance of the model without distribution effects. This is a flat distribution with segment values simply proportional to the row and column trip ends, written as t°.

8.5.4.1 Trip ends as fixed constraint

Because this initial model is fitted using screenline counts as well as the trip ends, its trip ends do not usually match those input as data. The initial model's trip ends will match the input trip ends if the input trip ends are taken as absolute constraints. In this case the initial model is completely defined by the trip end data as

$$t^\circ_{ij} = T_{i*} \times T_{*j} / T_{**}$$

and is independent of any screenline observation. The deviance of any screenline observation from this model is easily calculated, and the contributions of individual observations are independent and additive.

$$\text{Poisson deviance} = 2\Sigma(T\log(T/ t^\circ) -(T- t^\circ))$$

$$= 2\Sigma T(\log(T/ t^\circ) -(1- t^\circ/T))$$

$$= 2\Sigma T\, f(T/ t^\circ)$$

$$\text{where } f(x) = \log(x) -1+ 1/x \quad \text{is} \quad +\text{ve for } x>0$$

$$f'(x) = 1/x - 1/x^2 \quad \text{is} \quad -\text{ve for } x<1, +\text{ve for } x>1$$

The function $f()$ can be thought of as the amount of trip distribution information per screenline observation, T. It depends on the ratio x of observations to expectations from the fixed flat model based on trip end data, which in this case is also the null initial model against which the trip distribution model is being tested.

$$x_{ij} = T_{ij} \,/\, (T_{i*} \times T_{*j} \,/\, T_{**}) = T_{ij} \,/\, t°_{ij}$$

Two screenline observations from a trip distribution, T_a and T_b, with corresponding expectations from the flat model $t°_a$ and $t°_b$, might be available separately or in aggregate, $T_a + T_b$. The aggregate ratio can be re-written as a weighted average of the separate ratios

$$(T_a+T_b)/(t°_a+t°_b) \;=\; T_a/(t°_a + t°_b) \;+\; T_b/(t°_a + t°_b)$$

$$= \; T_a/\; t°_a \times t°_a \;/(t°_a + t°_b) \;+\; T_b/t°_b \times t°_b \;/(t°_a + t°_b)$$

The weighting is by the expected values from the flat model $t°$. These are positive for non-trivial cases, so the aggregate ratio must lie between the separate ratios, ie if

$$T_a/t°_a < T_b/t°_b$$

then

$$T_a/t°_a < (T_a + T_b)/(t°_a + t°_b) < T_b/t°_b$$

If the separate ratios are the same, then the aggregate ratio must be the same too. The function $f()$ is then the same for all three ratios:

$$f(T_a/\; t°_a) = f(T_b/\; t°_b) = f((T_a + T_b)/(t°_a+t°_b))$$

and so the sum of deviances of the separate observations is equal to the deviance of the aggregate observation.:

$$T_a\; f(T_a/t°_a) + T_b\; f(T_b/t°_b) = T_a f((T_a + T_b)/(t°_a+t°_b)) + T_b f((T_a + T_b)/(t°_a + t°_b))$$

$$= (T_a + T_b) f((T_a + T_b)/(t°_a + t°_b))$$

The function $f(T/\; t°)$ has a minimum of 0 where $T= t°$. If the ratios of the two observations to their expectations are opposed in the sense that $T_a< t°_a$ and $T_b> t°_b$ or vice versa, and hence lie to either side of the minimum, then the deviance can be reduced and possibly eliminated when the observations are aggregated. The function $f(T/\; t°)$ increases monotonically away from its minimum at $T/t° = 1$, so if the separate ratios both lie on the same side of unity the aggregate deviance must be greater than zero.

With typical trip distribution effects, intrazonal movements T_{11} and T_{22} will have lower costs and thus more trips than expected from the flat model, while interzonal movements T_{12} and T_{21} will have higher costs and fewer trips. If quadrants in one row or column are aggregated, their differences tend to cancel and the deviance can fall to zero. The resulting observations of a sector trip end total would not be expected to provide information about trip distribution within an existing set of trip ends as fixed constraints.

On the other hand, screenline crossings T_{12} and T_{21} will tend to be lower than expected from the flat model. Both must be lower (or higher) to match all the trip end constraints of the flat model. The ratios will be similar because of the broad symmetry in cost matrices. However, the ratios are also affected by the balancing factors, and since these are measures of accessibility, they depend on the spatial balance of productions and attractions and will not necessarily be symmetrical.

The ratios could be the same for the two quadrants. The total deviance would then be the same, and hence there would be no change in the amount of trip distribution information, if observations of the two PA quadrants were aggregated because they could not be distinguished in counts of traffic across the screenline. Otherwise, it has been shown for a general case that there will be some loss in deviance on

aggregation (section 3.7). However, the deviance cannot fall to zero because both ratios are less than 1. Hence the aggregate counts across a screenline will always provide some information about trip distribution.

This conclusion is for the special case where the trip distribution model is just determinate and the trip ends are taken as fixed constraints in the initial flat model. However, it appears to have more general application.

8.5.4.2 Screenline count as fixed constraint

Another special case is to take the screenline counts as absolute constraints in the initial flat model, ie

$$t°_{ij} = T_{ij}$$

where screenline counts are observed. This is indeterminate even if two quadrants are observed separately, so further criteria for the fit to input trip ends are needed. Because the initial flat model $t°$ is a function of screenline observations, the effects of screenline observations are no longer independent of each other as in the case of fixed trip ends.

If just one quadrant T_{11} is observed, the other quadrants in the initial flat model can be described by parameters ϕ_p and ϕ_a, where:

$$\phi_p = t°_{2*}/t°_{1*} = t°_{21}/t°_{11} = t°_{22}/t°_{12}$$

$$\phi_a = t°_{*2}/t°_{*1} = t°_{12}/t°_{11} = t°_{22}/t°_{21}$$

so the initial flat model $t°$ is:

Prod\attr	1	2	Trip ends
1	$\mathbf{T_{11}}$	$T_{11}\phi_a$	$T_{11}(1 + \phi_a)$
2	$T_{11}\phi_p$	$T_{11}\phi_p\phi_a$	$T_{11}\phi_p(1 + \phi_a)$
Trip ends	$T_{11}(1 + \phi_p)$	$T_{11}(1 + \phi_p)\phi_a$	$T_{11}(1 + \phi_p)(1 + \phi_a)$

Minimising the deviance of the input trip ends, T_{1*} etc from those in the model, $T_{11}(1+\phi_a)$ etc gives cubics in ϕ:

$$\begin{aligned}
&\phi_p^3 (-2T_{11}T_{2*}) \\
+\ &\phi_p^2 (-4T_{11}T_{2*} + T_{**}T_{2*} - T_{**}T_{2*} + T^2_{*2} - T_{2*}T_{*2}) \\
+\ &\phi_p (-2T_{11}T_{2*} + T_{**}T_{2*} + 2T^2_{*2} - T_{2*}T_{*2}) \\
+\ &T^2_{*2} \\
=\ &0
\end{aligned}$$

The cubic in ϕ_a is given by transposing the ij indices.

These are solved and the residual deviances are shown in figure 8.2 for a nominal case of a unit screenline count, $T_{11} = 1$, and equal productions and attractions in each sector. These trip ends are set at $2/x$, so the flat matrix derived from them alone has an expectation of $1/x$ in each cell. The unit screenline count is thus x times the expectation from the trip ends without trip distribution effects, as in section 8.5.4.1, but this expectation is no longer the null initial model against which the trip distribution model is compared.

$$x_{ij} = T_{ij} / (T_{i*} \times T_{*j} / T_{**}) \neq T_{ij} / t°_{ij}$$

From symmetry, $\phi_p = \phi_a = \phi$, whose maximum likelihood condition reduces to

$$x\phi^3 + 2x\phi^2 + (x-3)\phi - 1 = 0$$

$$\text{or} \quad x(1+\phi)^2\phi = 3\phi + 1$$

Figure 8.2 Deviance with absolute constraints

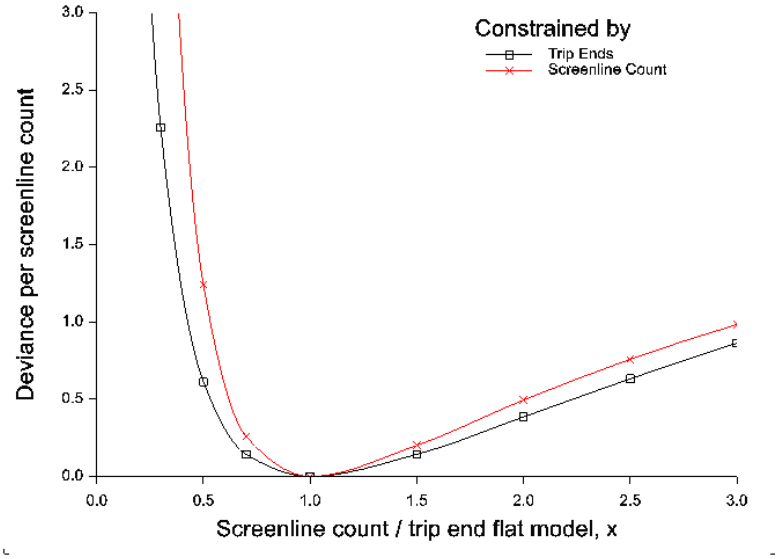


Figure 8.2 also shows the deviance for the equivalent case constrained by trip ends; this is twice the function $f(x)$ considered in section 8.5.4.1.

Both functions have a minimum of 0 at x = 1, where the screenline count is consistent with the flat model defined by the trip ends alone and cannot indicate any trip distribution effect.

There is a practical limit at x = 2, where the screenline count for one segment is equal to the whole of the trip ends for the corresponding sectors. Larger values of x demand negative trips in other segments to comply with both trip end and screenline data in a trip distribution model.

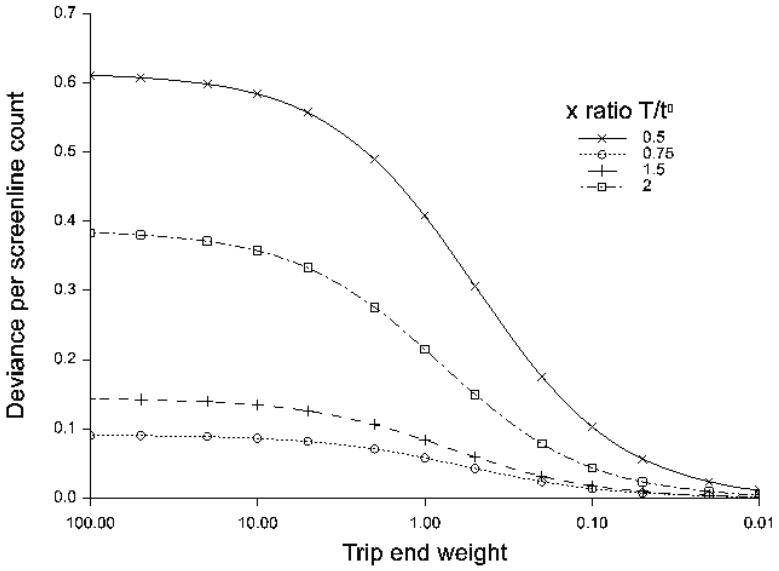
The lower deviances for x greater than unity are in part due to the reduction in trip end data to maintain the ratio x with the unit screenline count. As x becomes small, trip ends become large

8.5.4.3 Relative weighting of screenlines and trip ends

Between the two limiting cases of fixed constraints considered above, there is a continuum of relative weighting between screenline count and trip end data. Relative weighting is readily accommodated by the probabilistic approach of MVESTM and GLMs, unlike absolute constraints which are better handled by methods of optimisation and mathematical programming.

Figure 8.3 plots deviances against relative weighting for selected values of x, the ratio of a screenline count of a single segment to its expectation from trip end data without trip distribution effects. All trip end data are equal and the screenline count is set at unity.

Figure 8.3 Deviances with varying weights



The weighting of the screenline count is also fixed at unity (or 100 in terms of MVESTM confidence), while the trip end weighting varies along the horizontal axis.

Strong trip ends are plotted to the left and weak ones to the right, corresponding with the layout of later tables. The left side approximates to trip ends as absolute constraints, with values similar to those in figure 8.2.

The flat ends of the reverse S shapes show that little trip distribution information is available without sufficient trip end data and that there is a limit to the trip distribution information that can be gained by strengthening the trip end data alone. In between, the amount of trip distribution information is more sensitive to the strength of trip end data.

The range of relative weights corresponds with sensitivity tests on the household data. Since both trip end and screenline data was derived from the same fully observed matrix, their ordinary weighting was equal. For strong or weak trip ends, their weights were multiplied or divided by a hundred. In the practical case of actual screenline counts and trip ends from a generation model, ordinary trip end confidences were set at about 0.1 of the screenline confidences from the limited evidence discussed in section 8.4.4.

The figure could be replotted with a fixed trip end weight and varying screenline weights. The deviances would then fall towards zero for small screenline weights – both trip end and screenline count data are needed to show trip distribution effects – and approximate to the case of screenline counts as absolute constraints for large weights.

8.6 Data sources and preparation

The ultimate objective was to demonstrate the calibration of trip distribution from traffic counts without recourse to a fully observed trip matrix. However, the analyses started with a fully observed trip matrix to demonstrate consistency with conventional calibration that requires such disaggregate data.

To allow comparisons with previous analyses and to avoid the complications of multiple purposes (discussed later in section 8.8.1.3), calibrations from both full matrices and from screenline counts were

of distributions of internal HBW person trips by car. Adjustments to observed traffic counts using the WTSM model are described in section 8.6.4.

The WTSM model also provided the proportion of each OD movement intercepted at each screenline, and period, direction and occupancy factors.

8.6.1 Full trip matrix

The full matrix was observed in the WTSM Household Interview Survey (HIS). It was a PA matrix for a 24-hour weekday, as had been used for previous calibrations by GLM.

The synthetic matrices from the GLM calibrations were also analysed, since it was known that trip distribution models could fit these matrices exactly. There was one synthetic matrix for each of three deterrence functions – Exponential, Tanner or Power.

Aggregations of these full matrices follow the sector systems of spatial analysis in chapter 7, figure 7.2 with sets of 65, 15, 10, 6, or 3 sectors.

The zone-to-zone movements comprising each of the resulting segments are simply derived from the production and attraction sectors. This information is needed in the intercept file to define the scope of each aggregate trip total.

8.6.2 Trip ends

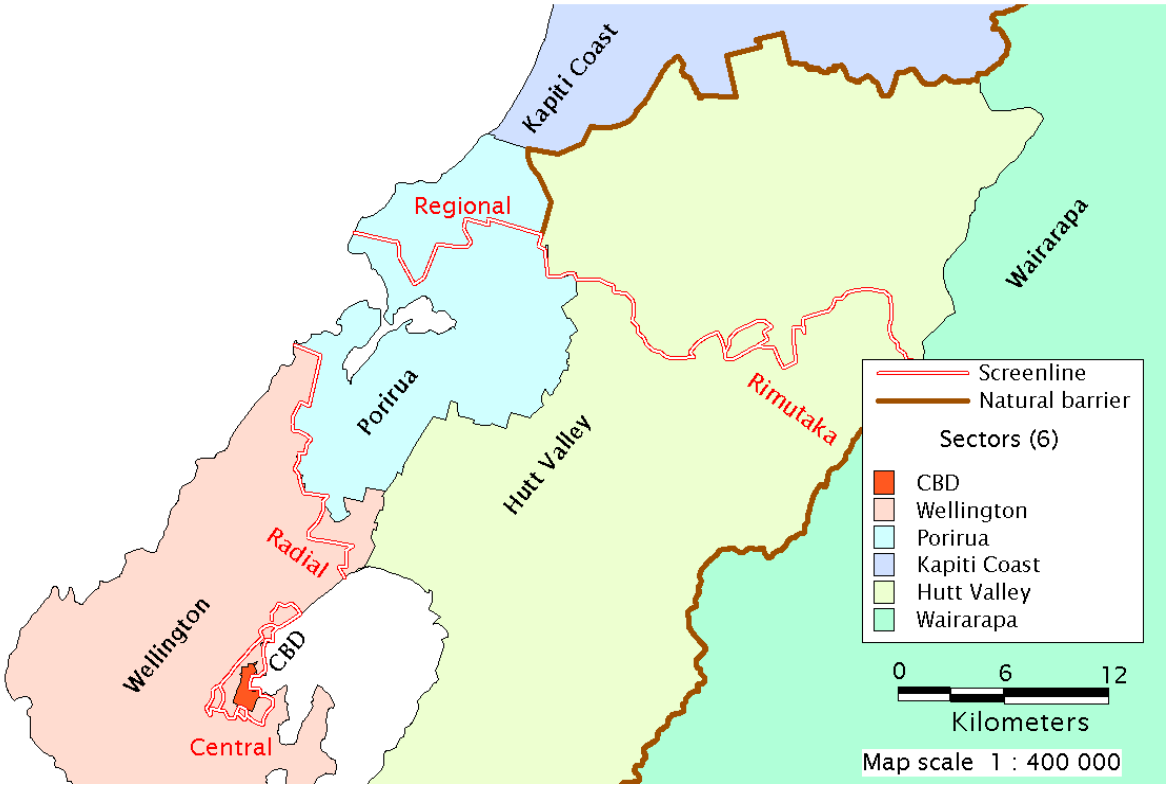
Any trip ends required to complement the full HIS matrix were simply taken from its own row and column totals. The synthetic matrices conform to the same set of trip ends as the observed matrix.

The full HIS matrix was avoided as a source for calibration from traffic counts. In practice, the likely source of trip ends in the absence of a fully observed matrix is a trip generation model based on planning data. Synthetic trip ends from the WTSM production and attraction models were used to complement traffic counts at screenlines. The trip ends were actually totalled from the synthesised base year HBW car matrices (by three household car availabilities) after the joint distribution and mode split in the WTSM, to give 24-hour PA person trips by car.

8.6.3 Screenline counts

8.6.3.1 Screenline location

Figure 8.4 Screenlines and sectors



The screenlines correspond broadly with three- and six-way sector boundaries. The radial screenline separates Wellington city from the other two of the three-way sectors, SH1 Kapiti and SH2 Hutt. This similarity is used in the transition from the fully observed HIS matrix to the actual screenline counts in section 8.7.6.

The central, regional and Rimutaka screenlines split the city, Kapiti and Hutt three-way sectors roughly into the six-way sectors. Exclaves, where screenlines differ from natural boundaries, are shown and discussed in section 8.7.7.7.1.

The central screenline is a cordon surrounding Wellington city centre, so there are double crossings between south Wellington and north Wellington, Kapiti and Hutt; some movements between these areas and parts of west Wellington are also intercepted twice. North, south and west Wellington are distinguished in the 10-way sectors.

8.6.3.2 Intercepts

The scope of movements intercepted by each screenline, and thus contributing to the traffic count, has to be defined in the intercept file.

Unlike the aggregations of the full HIS matrix which are simply defined by sets of production and attraction zones, the movements intercepted by the actual screenlines sometimes depend on routings through the network. Where there are alternative routings, such as around or through the central cordon, only a proportion of a movement might be intercepted but it might be counted more than once.

The WTSM model provided the expected count of OD movements across each screenline in each direction. Select link matrices were produced as an additional option in the EMME/2 base year assignments for each of the three periods, AM, IP and PM. Table 8.6 shows that the number of movements intercepted at the central and radial screenlines varied slightly between the three periods because of congestion effects. With no alternative routes around the regional and Rimutaka screenlines, the number of intercepted movements is the same for all periods and directions. The derivation of the intercepts and some of the routing issues are described in appendix C.

Table 8.6 Number of OD movements intercepted at screenlines

Screenline	AM in	AM out	IP in	IP out	PM in	PM out
Central	26,358	24,784	25,548	26,244	24,780	25,570
Radial	27,402	25,818	26,482	27,402	25,818	26,482
Regional	8200	8200	8200	8200	8200	8200
Rimutaka	10,000	10,000	10,000	10,000	10,000	10,000

8.6.3.3 Period, direction and occupancy

The conversion from an all-day production–attraction trip distribution to hourly directional counts is incorporated into the intercept file. For each zone-to-zone movement in the intercept file, there is a probability R_{ijk} of the movement being observed at a screenline. This is conventionally used for multi-route assignment, so if only a quarter of trips from i to j are intercepted at screenline k, $R_{ijk} = 0.25$.

This concept is extended so that if an hourly flow during a certain period is only one-tenth of the daily total, the intercept proportion R_{ijk} incorporates a factor of 0.1 for the hourly counts from that period.

Vehicles were counted by the direction in which they are travelling. A directional count comprises both home-to-work trips, going from production to attraction, and work-to-home trips, going from attraction to production, but cannot distinguish them. The select link matrix is defined by direction of travel from origin to destination, but the intercept file is applied to the 24-hour production-attraction matrix. Intercept proportions for the home-to-work trips are calculated simply from the select link matrix multiplied by the appropriate direction and period factors. To these are added the intercept proportions for the work-to-home trips, which are the transpose of the select link matrix multiplied by different direction and period factors.

The intercept file also incorporates occupancy factors to convert from person trips by car (including passenger) to vehicle trips.

Period, direction and occupancy factors were taken from the WTSM model. Their product is summarised in table 8.7.

Table 8.7 WTSM 24-hour PA person to hourly OD vehicle trip factors

Movement	Home to work (OD=PA)			Work to home (OD=AP)		
	AM	IP	PM	AM	IP	PM
Wellington TLA <> Wellington TLA	0.143	0.014	0.007	0.004	0.010	0.085
Wellington TLA <> other	0.149	0.014	0.007	0.004	0.006	0.097
Other <> other	0.140	0.014	0.007	0.004	0.015	0.085

Source TN 19.1; HBW
AP = attraction to production

Period and occupancy factors vary by segments divided between Wellington territorial local authority (TLA) and other internal areas. Wellington city TLA extends beyond Johnsonville to include Tawa, unlike the Wellington city sector of the three sectors used in aggregations of the full HIS matrix. This segmentation of factors is symmetric, so the factors can be applied either before or after the transposition of the select link matrix for work-to-home trips.

8.6.4 Counts

Screenline counts were taken from the classified counts of cars and light commercial vehicles used for validating the WTSM. These were found on the 'ObsVol' sheet of the spreadsheet slvalid.xls. Counts from this sheet appear in tables 3-2 to 3-4 of TN22.1 'Validation report'.

8.6.4.1 Adjustments to observed counts

The trip distribution is of internal home-based work trips. The vehicle counts at screenlines cannot distinguish trip purpose or external trips. The counts have been factored down from all modelled purposes to HBW, excluding airport and external trips. The proportions of internal HBW to all car trips were taken from assignments of the appropriate synthesised WTSM base year matrices to each screenline, by period and direction. Select link matrices were applied to find screenline crossings under a consistent set of routings. The resulting proportions are shown in table 8.8.

Table 8.8 Percentage of all car trips that are internal HBW

Screenline	AM in	AM out	IP in	IP out	PM in	PM out
Central	63.6%	35.0%	13.2%	10.3%	21.0%	42.7%
Radial	63.1%	46.0%	13.7%	10.5%	27.6%	43.3%
Regional	57.4%	24.7%	11.7%	9.2%	14.9%	37.9%
Rimutaka	53.7%	30.2%	10.5%	9.9%	16.5%	31.4%

From assignments of WTSM base year matrices

The resulting hourly counts by screenline, period and direction are shown in table 8.30.

8.6.4.2 Other features in the WTSM

The WTSM introduces other sub-models between the trip distribution and final assignment stages which are not represented in the intercept factoring or count adjustment described above.

Buses and heavy commercial vehicles were excluded both from the classified counts of light vehicles and from the person-trip modelling, which included light commercial vehicles.

Differential growth of light commercial vehicles and peak spreading are pivoted about the base year and so are null in the base year model.

Matrix adjustments (from matrix estimation with MVESTM) and further adjustments of 2.5% for AM and PM were applied to improve the fit of the synthetic model to screenline counts. Since this is the fit being investigated here the adjustments have been omitted.

8.6.5 Costs

Costs are the generalised cost used in the WTSM for HBW car trip distribution, but with parking charges omitted and intrazonal costs, based on half the minimum interzonal cost, amended accordingly.

This definition was adopted in part for consistency in chapter 7 'Spatial patterns', where costs are used as a measure of spatial separation as well as a deterrence to travel. It was hoped this would help to recognise any complementarity between the two methods in analysing hierarchical aggregations.

8.7 Analyses

The data analyses are intended to demonstrate:

- first, consistency with other methods of trip distribution modelling
- then, calibration from screenline counts, in a practical case.

The initial demonstration of consistency used a fully observed trip matrix, as is normal for conventional calibration. Calibration by MVESTM encountered computational limitations, which were overcome by aggregating the data. This gave insights into the properties of aggregate data, which is at the heart of MVESTM and estimation from screenline data. It also opened up a pathway from the observed HIS trip matrix, already analysed extensively by GLM, to traffic counts by period and direction as the data source for calibrating a trip distribution model that links productions with attractions. This was followed step-by-step through the following stages:

- calibration from all segments aggregated from the HIS matrix, at different levels of aggregation
- calibration from each single segment of the HIS matrix aggregated to 3 x 3 segments
- calibration from quadrants and diagonal pairs of quadrants of the HIS matrix, in a further aggregation from three to two sectors
- transition from the HIS matrix to screenline counts based on the similarity between the boundary between the two sectors and the radial screenline, and between the trip totals for the two intersector quadrants of a 24-hour PA trip matrix and the all-day, two-way counts of traffic crossing the screenline
- calibrations on counts from four screenlines, in three periods and two directions, individually or in combination. Combinations of counts are entered either as separate items of data or as a single aggregate count.

The aggregation of the HIS matrix also brought in trip end data. This was originally introduced to avoid indeterminacy in the estimated matrix. It was not expected to affect the calibration of cost parameters, but a small effect was found. The need for trip end data is more obvious in calibration from single matrix segments and is also apparent in calibration from screenlines. A theoretical basis for the need for trip end information is set out in section 8.5.

Early stages of the development started with synthesised trip distributions as inputs, where the deterrence function was known and an exact fit was possible. Count data was synthesised from the matrices. Once the fit in these ideal cases was established, the methods were applied to observed trip distributions and counts. The very first stage was to synthesise such models using MVESTM itself.

8.7.1 Synthesis

Synthesis is a simpler task than calibration, using the cost parameters already calibrated by a GLM to replicate the distribution it fitted.

8.7.1.1 MVESTM formulation

Table 8.9 MVESTM formulation for synthesis

Data	Scope	Structure
Prior matrix		
Costs		$t_{ij} = c_{ij}^{-\gamma} \exp(-\lambda c_{ij})$
Trip ends		**Parameters**
Totals from trip matrix		Free to fit to trip end total x balancing factor
Screenlines	**Intercepts**	
No file	*No file*	*none*
		Alpha, beta fixed at calibrated values of $-\gamma$, λ

Italic – null effect on fitted model

Note: cf table 8.2 or E.1

The formulation for MVESTM to synthesise a trip distribution is relatively straightforward. Costs are entered through the prior matrix and trip ends are entered through the trip end file. No screenline or intercept files are needed, but a parameter file has to be prepared to include the cost coefficients of the deterrence function as fixed parameters. The trip end parameters are left free, and have to represent both balancing factors and the trip end total.

8.7.1.2 Fit

Table 8.10 shows differences between the trip distribution synthesised by MVESTM and the distribution produced during calibration by GLM.

Table 8.10 Differences between GLM calibration and MVESTM synthesis

Deterrence function	Difference in trips by matrix cell		Deviance
	Minimum	Maximum	
	Trip end parameters **not** initialised		
Exponential	-0.0190	0.0039	3.15×10^{-6}
Tanner	-0.0461	0.0024	11.88×10^{-6}
Power	-3.4181	0.4754	69602.13×10^{-6}
	Trip end parameters initialised		
Exponential	-0.0188	0.0182	3.98×10^{-6}
Tanner	-0.0029	0.0031	0.08×10^{-6}
Power	-0.0006	0.0083	0.16×10^{-6}

Differences are given without and with trip end parameters initialised, as described in section 8.3.6. Without the parameters initialised, there is a case of non-convergence for the Power function. With a greatest difference of just 3.4 trips, it is negligible in practical terms, but quite distinct in this analysis. With the trip end parameters initialised, this non-convergence is eliminated.

The deviances are for the fit of the synthesised MVESTM 'estimates' to 'observations' of trip distributions calibrated by GLM. They include weighting of 1/157.9 used in other calibrations of HIS data for

comparison. They show that the effects of computation and convergence in the deviance are small and can be very small.

8.7.2 Calibration on full, disaggregate PA matrix

8.7.2.1 MVESTM formulation

Table 8.11 MVESTM formulation for calibration on full, disaggregate PA matrix

Data	Scope	Structure
Prior matrix		
Costs		$t_{ij} = c_{ij}^{-\gamma}\exp(-\lambda c_{ij})$
Trip ends		**Parameters**
No file		Free to fit to trip end total x balancing factor
Screenlines	**Intercepts**	
All PA trip data	One ij cell per screenline, all i,j	*Fixed at unity*
		Alpha, beta free to calibrate $-\gamma$, λ

Italic – null effect on fitted model

Note: cf table 8.2 or E.1

Conventional trip distribution calibration is more difficult to formulate in MVESTM. It is not possible to enter both costs and observed trips through the prior matrices. Costs cannot be entered in any other way, so the observed trips are entered as screenline data with every production–attraction pair treated as a screenline. An intercept file is prepared to represent this structure. The corresponding screenline parameters are fixed. The cost parameters are left free for calibration, and the trip end parameters are free to fit the balancing factors and trip end totals. No trip end file is entered, since the totals are implicit in the prior trip observations entered as screenlines.

8.7.2.2 Limits on the number of screenlines

Representing each cell of the observed trip matrix as a separate screenline required 31,428 screenlines after excluding empty zones. This presented problems of numbering and of core memory allocation which could not be resolved (see appendix E, section E.2.4). Instead, the observed trip data was aggregated from individual cells to a smaller number of segments.

Later in the study, trip data from the 31,428 individual matrix cells was successfully presented to MVESTM as part-route (level 2) counts in the volume fields of a network. See appendix E, section E.1.5.

8.7.3 Calibration on full, aggregated PA matrix

8.7.3.1 Aggregation by segment

The segments are defined by the same sets of 3, 6, 10, 15, or 65 sectors that are used in chapter 7 'Spatial patterns'. One screenline count presents the total observed matrix trips for a whole segment, instead of a single zone-to-zone movement, reducing the number of screenlines.

Segmentation to 65×65 sectors fits into available memory.

8.7.3.2 MVESTM formulation

Table 8.12 MVESTM formulation for calibration on full, aggregated PA matrix

Data	Scope	Structure
Prior matrix		
Costs		$t_{ij} = c_{ij}^{-\gamma}\exp(-\lambda c_{ij})$
Trip ends		**Parameters**
Totals from trip matrix, to provide proportions within sectors		Free to fit to trip end total x balancing factor
Screenlines	**Intercepts**	
All PA trip data aggregated to segments	One segment per screenline, all segments	*Fixed at unity*
		Alpha, beta free to calibrate $-\gamma$, λ

Italic - null effect on fitted model

Note: cf table 8.2 or E.1

The formulation is essentially the same as for disaggregated PA data, but with a smaller number of screenlines, and the intercept file specifying the sets of zone-to-zone movements that comprise each segment. However, a trip end file is added to determine the allocation of trips within each sector.

8.7.3.3 Inclusion of zonal trip end data

Without this zonal trip end information, there is no clear basis for the distribution of trips between zones within a sector. When prior trips are presented cell by cell, their summation by origin or destination gives an observed trip end total for each zone. When they are aggregated by segment, they only provide trip end totals by sector and further information is needed to distribute trips between zones within the sector. There is redundancy of information about the sector trip ends, but no differences if the zonal trip ends are calculated from the same prior matrix.

It was expected that trip end data would provide determinacy in the model without affecting the calibration of the cost coefficients and that the estimated zonal trip ends would match the observations exactly. When it was found that these were not strictly so, the influence of the trip end data was investigated by introducing it at three levels of confidence: strong, ordinary and weak.

The ordinary level of confidence is the same as that for the segment trip totals (100/157.9, based on the expansion from observed home-workplace pairings to trips). The strong confidence is 100 times greater, approximating the trip end data to a fixed constraint. The weak confidence is 100 times less than the ordinary, leaving the trip end data subsidiary to the segment totals.

8.7.3.4 Fit to synthetic data

Models were first calibrated on synthetic data, in the form of trip distributions with the appropriate deterrence function fitted to the observed data by GLM. This data can be fitted exactly by the MVESTM models.

MVESTM achieves a good fit to this synthetic data, with total residual deviances of 0.01 or less, compared with the fit to observations in table 8.14, 5th column. These residuals include the fit at zonal trip ends (cf table 8.18 for fit to observed data). Cost coefficients are replicated to within one part per thousand, and often much closer, of those used to synthesise the data (cf table 8.17). The one exception is Tanner models aggregated to 3x3, which have only two residual degrees of freedom after fitting the two correlated coefficients; the worst error in these coefficients is 9%.

Given these good fits, variations with aggregation in the following estimations from observed data reflect imperfections in models' fit to observations, rather than artefacts of the methodology or its computation.

8.7.3.5 Fit to observed data

There will be a lack of fit to observed data due to sampling error. If this follows a Poisson process, the total residual deviance is expected to approximate to a χ^2 distribution with the appropriate degrees of freedom.

Table 8.13 Degrees of freedom and sparsity

Aggregation	Non-empty sectors			Degrees of freedom			Effective observations per segment	Expected mean deviance
	Production	Attraction	Segments	Trip ends	Common	Combined		
3x3	3	3	9	356	6	359	128.93	1.001
6x6	6	6	36	356	12	380	32.23	1.005
10x10	10	10	100	356	20	436	11.60	1.016
15x15	15	15	225	356	30	551	5.16	1.045
65x65	**58**	65	3770	356	123	4003	0.31	0.849
Zonal	**162**	**194**	31428	356	356	31428	0.04	0.246

Bold – reductions due to empty sectors or zones

In table 8.13 the degrees of freedom are calculated from the effective number of sectors or zones, excluding empty ones with no observations. The number of segments is the product of the numbers of production and attraction sectors, and the number of zonal trip ends is the sum of production and attraction zones. The sector trip ends are common to both data sets and are therefore subtracted from the sum of their degrees of freedom for the combined data set.

Fitting even a null flat model, with matrix cells simply proportional to trip ends, requires 355 parameters – one for an overall mean plus one less than the numbers of both production and attraction zones. These can be fitted from the trip end data alone, leaving segment data to represent trip distribution effects, and determine how well the flat model fits them. Without trip end data, 15 x 15 or fewer segments have insufficient degrees of freedom and are clearly indeterminate. The 65 x 65 segments do offer enough degrees of freedom to avoid simple indeterminacy, but are structured so that individual trip ends cannot be determined directly from the segment totals.

The approximation to the χ^2 distribution becomes poor as observations become sparse (see sections 3.4 and 3.5). The last two columns of table 8.13 show the average number of observations per segment and the corresponding expected mean deviance, which is unity for large numbers of observation. The effective total number of observations, 1160, is the expanded trips, 183,216, divided by the overall expansion factor and inverse weight, 157.9. This takes one worker's home-workplace pairing as the unit of observation. On average, segmentation down to 15 x 15 does not make the observations sparse and the expected mean deviance for the average is close to unity. However, 65 x 65 and zonal segmentations are sparse, with expected mean deviances well below unity. Average zonal trip end observations are 7.2 for productions and 6.0 for attractions – not sparse.

However, trips are not usually distributed equally between segments. Table 8.14 shows the residual deviances for a null model and for a trip distribution model estimated on observed data, together with the mean deviances expected from the fitted models. The residual degrees of freedom are the combined degrees of freedom from table 8.13 less 355 for fitting trip end parameters in the null model; and one less for fitting the cost coefficient in the distribution model.

Table 8.14 Fitted and expected residual deviances

Aggregation	Null, flat model				Estimated trip distribution model (Exponential, ordinary trip end confidence)			
	Total	df	Mean	Expected	Total	df	Mean	Expected
3 x 3	959.5	4	239.87	1.002	3.0	3	1.02	1.006
6 x 6	1661.9	25	66.48	0.994	47.1	24	1.97	0.802
10 x 10	1979.7	81	24.44	1.006	112.0	80	1.40	0.718
15 x 15	2136.4	196	10.90	1.034	197.3	195	1.01	0.706
65 x 65	3641.2	3648	1.00	0.621	1543.7	3647	0.42	0.418
Zonal	6380.4	31073	0.21	0.191	4249.5	31072	0.13	0.140

Under the null flat model, without cost deterrence effects, the variations in segment totals are simply due to uneven distribution of productions and attractions between sectors. Zonal and 65 x65 segmentations show increased effects of sparsity with reductions in the expected deviance. Greater aggregations show no such effects and here the fitted mean residual deviances clearly show a lack of fit to observations.

This lack of fit is far less clear at the zonal level. The small excess of the mean residual over its expected value may be significant for the large degrees of freedom but is difficult to test.

Introducing the cost deterrence effects of trip distribution reduces the number of trips fitted in many segments. Sparsity effects become apparent in the expected deviance at all levels of aggregation except 3x3. At the zonal level there is under-dispersal of the mean residual deviance compared with expectation and at other levels there is no longer the same gross lack of fit that is apparent in the flat model.

- Disaggregation to zonal level may disguise poor fit; it becomes harder to assess from residual deviances with sparsity.
- Although the mean residual deviance from the flat model at zonal level 0.21 looks reasonably close to its expected value 0.191, it looks much worse against the expected deviance of the trip distribution model 0.140.
- Although there is little sign of misfit to the Exponential trip distribution model at many levels of aggregation, changes in deviance (table 8.15) show that the Tanner deterrence function can improve on it very significantly at most levels.

The estimated trip distribution in table 8.14 has an Exponential deterrence function and ordinary trip end confidences. It is given as one example of residual deviances.

8.7.3.6 Information about trip distribution

Table 8.15 shows the reduction in deviance of every estimated trip distribution, compared with a null, flat model. These changes in deviance represent the significance of cost in trip distribution, which is huge in every case. The first three columns show the fit to observed data with different confidences in trip ends; the final column shows the fit to synthetic data.

Table 8.15 Change in deviance for all aggregate HIS segments

Aggregation	Trip end confidence			Synthetic data
	Strong	Ordinary	Weak	
	Exponential deterrence function			
3 x 3	956.2	956.4	959.0	902.7
6 x 6	1614.1	1614.8	1633.6	1670.2
10 x 10	1865.8	1867.7	1904.6	1924.9
15 x 15	1937.0	1939.1	1976.5	1989.2
65 x 65	2096.3	2097.6	2132.2	2100.0
Zonal (by GLM)	*2130.9*	*2130.9*	*2130.9*	*2130.9*
	Tanner deterrence function			
3 x 3	956.8	957.6	959.0	883.6
6 x 6	1640.4	1640.8	1651.1	1604.5
10 x 10	1917.5	1918.2	1932.6	1890.6
15 x 15	1990.5	1991.7	2019.4	1962.4
65 x 65	2153.1	2155.5	2222.5	2126.7
Zonal (by GLM)	*2197.0*	*2197.0*	*2197.0*	*2197.0*
	Power deterrence function			
3 x 3	949.0	949.7	958.4	709.5
6 x 6	1619.4	1621.9	1649.7	1187.3
10 x 10	1881.3	1883.7	1927.9	1492.2
15 x 15	1934.1	1938.6	2012.1	1579.6
65 x 65	2021.8	2029.7	2168.8	1853.6
Zonal (by GLM)	*2027.5*	*2027.5*	*2027.5*	*2027.5*

italic – fitted by GLM, without explicit trip end data.

In a simple experimental analysis, the change in deviance is proportional to the amount of data collected. In each sub-column of this table, all the data is derived from the same observed matrix and the change in deviance may be interpreted as the amount of information about trip distribution that is abstracted at different levels of aggregation.

Under this interpretation, only about a quarter of information is lost in aggregation from 225 zones up to 6x6 segments. More information is lost in the final aggregation from there to 3 x 3 segments, leaving about 40% of the original from the zonal level.

8.7.3.6.1 Zonal and 65-sector aggregations of Power models

Surprisingly little information is lost in the initial aggregation from zones to 65 x 65 segments, particularly for Power models. As trip end constraints are loosened, towards the right of the table, distribution effects appear more significant in the 65 x 65 aggregations than in the zonal GLM models.

This appears to contradict the expectation that information will be lost in aggregation. The workings in section 3.7 that support this expectation apply to the change in deviance from a mis-specified model to a true one. While the flat initial model is clearly mis-specified, the trip distribution models are not necessarily true.

The component deviances are set out in table 8.16 for the most incongruous case, of the Power model with weak trip ends. Two columns show deviances assessed either at the 65 x 65 aggregation or at the zonal level. At the aggregate level, the deviances of zonal trip ends are included as they are part of the model formulation and hence objective function fitted at that level.

Table 8.16 Deviances for Power models with weak trip ends, zonal vs 65 sector aggregation

Model fitted to	Deviance calculated for	
	65x65 segments + zonal trip ends	Zone x zone
	Residual deviances	
Zonal trip ends – initial, flat model	3641.2 + 0	6380.4
65 × 65 segments & zonal trip ends	1445.3 + 27.2	7243.1
Zone × zone	1623.0 + 0.00000041	4352.9
	Change in deviances	
65 × 65 segments & zonal trip ends	2168.8	-862.6
Zone × zone	2018.3	2027.5

The top three rows show residual deviances for three models

1 The initial null model, which is a flat model without a cost component, with matrix cells simply proportional to zonal trip ends

2 The trip distribution model fitted to 65 x 65 segments, and zonal trip ends

3 The trip distribution model fitted to the zonal matrix. Zonal trip ends are implicit in the row and column totals.

These deviances are always smaller when calculated at the aggregate level, in the first column.

At both levels of calculation, the trip distribution model fitted at that level has the lower deviance as would be expected from the objective of minimising the deviance.

The model fitted to 65 x 65 segments fits remarkably badly at the zonal level. At that level, the fit is worse even than the flat initial model with no cost deterrence effects. Given there is a fair fit to segment totals and to zonal trip ends, it would seem these are achieved by some arrangement of zone-to-zone trips within each segment that is very much at odds with the observed trips.

This appears to be facilitated by adjustment of the weak trip ends, since the effect is much smaller with stronger trip ends.

The effect is most marked for the Power deterrence and may be another aspect of this function's sensitivity to intrazonal costs, noted in section 4.11.

The bottom two rows show changes in deviance from the initial flat model as the cost term is introduced to give a trip distribution model. The model fitted at the zonal level shows little reduction in this change of deviance when re-assessed at the aggregate level, in stark contrast with the distribution model fitted at the aggregate level.

8.7.3.6.2 Trip end confidences

In table 8.15 there are differences in the change of deviance according to the confidence levels of trip ends. However, they are small considering the large variations in weighting attached to trip ends, suggesting a weak interaction between cost effects and within-sector trip end constraints.

8.7.3.6.3 Synthetic data

With synthetic data, matrix estimation can and does achieve an exact fit of the distribution models, with no residual deviance. The changes in deviance shown in the final column of table 8.15 therefore arise from fit of the null, flat model to the synthesised trip distributions and are actually the deviances from that flat model at different levels of aggregation.

They show the amount of information that can be recovered at different levels of aggregation from a 'pure' trip distribution model, omitting random sampling and any systematic departure from such models in the observed data. Very broadly, they follow the patterns for observed data. However, there is less information about the Power model from aggregate data, suggesting that the observed data carries information that supports the Power model in an upper hierarchy between segments.

8.7.3.6.4 Deterrence functions

Exponential deterrence functions fit observations distinctly better than the Power functions at the zonal level. This is reversed for 6 x 6 and 10 x 10 aggregations with strong or ordinary trip end confidences, and at all levels of aggregation except 3 x 3 with weak confidences. Much evidence against the Power function appears to lie in the within-sector distribution of zonal trip ends, despite their weak interaction with distribution effects.

There is generally strong evidence for the improved fit to observations of the Tanner deterrence function over either the Exponential or Power functions. There is little evidence of improvement over the Exponential at 3 x 3 aggregation, or over the Power for 3 x 3 or 6 x 6 aggregations and weak trip end confidence. Because the Power and Exponential functions are sub-models within the Tanner function, the Tanner function's improvement can be tested formally, unlike the differences between the Power and the Exponential functions which are not nested.

A different dataset is synthesised for each deterrence function, so the deterrence functions cannot be compared directly from the deviance changes in the final column of table 8.15. It is notable that at higher levels of aggregation more distribution information is retrieved for the Exponential function than for the Tanner function, which would not be possible in nested models.

8.7.3.7 Fitted cost coefficients

Table 8.17 Cost coefficients fitted to aggregated HIS segments

Aggregation	Trip end confidence					
	Strong		Ordinary		Weak	
	λ, Cost	γ, LnCost	λ, Cost	γ, LnCost	λ, Cost	γ, LnCost
	Exponential deterrence function					
3 x 3	0.068	~	0.069	~	0.073	~
6 x 6	0.060	~	0.060	~	0.060	~
10 x 10	0.061	~	0.061	~	0.056	~
15 x 15	0.061	~	0.062	~	0.059	~
65 x 65	0.064	~	0.064	~	0.064	~
Zonal	*0.064*	~	*0.064*	~	*0.064*	~
	Tanner deterrence function					
3 x 3	0.087	-0.46	0.148	-2.02	0.066	0.17
6 x 6	0.022	1.09	0.022	1.10	0.012	1.50
10 x 10	0.023	1.03	0.023	1.03	0.020	1.09
15 x 15	0.027	0.91	0.027	0.92	0.016	1.32
65 x 65	0.035	0.70	0.034	0.71	0.026	1.01
Zonal	*0.036*	*0.65*	*0.036*	*0.65*	*0.036*	*0.65*
	Power deterrence function					
3 x 3	~	1.71	~	1.72	~	1.99
6 x 6	~	1.77	~	1.78	~	1.95
10 x 10	~	1.64	~	1.65	~	1.91
15 x 15	~	1.60	~	1.61	~	1.95
65 x 65	~	1.46	~	1.48	~	1.69
Zonal	~	*1.40*	~	*1.40*	~	*1.40*

Italic – fitted by GLM

There is a general trend for Exponential coefficients λ to diminish and Power coefficients γ to increase with increasing aggregation up the tables. Bly et al (2001, end of section 8.3) found lower Exponential coefficients in (only three) regional models than in urban models. The differences here appear at different levels of aggregation of the same set of observations.

The trend is often reversed in the final aggregation to 3 x 3 segments. The aggregation process is not a uniform one, and this final stage merges the primary attractor of the whole study area, Wellington city centre, with its immediate hinterland. This is not ideal zoning for a distribution model and table 8.15 shows a considerable loss of information. At this 3 x 3 aggregation, Tanner models for strong and ordinary trip end confidences are fitted with positive Power coefficients. This gives a convex, humped form to the deterrence function, rather than the continuously diminishing concave form fitted in all other cases.

The general patterns of variation with aggregation are similar for all trip end confidences, but the extent of variation increases as the confidence diminishes towards the right of the tables.

8.7.3.8 Fit at trip ends

Mismatches appear at zonal trip ends when estimating trip distributions from observed data. Estimations of the flat model or from synthetic data match well, with total residual deviances typically of the order 10^{-6} or less.

Table 8.18 Zonal trip end deviances

Aggregation	Trip end confidence		
	Strong	Ordinary	Weak
	Exponential deterrence function		
3 x 3	0.002	0.19	0.40
6 x 6	0.008	0.78	5.86
10 x 10	0.052	1.66	13.66
15 x 15	0.053	1.66	13.88
65 x 65	0.018	1.20	14.32
	Tanner deterrence function		
3 x 3	0.002	0.22	0.45
6 x 6	0.005	0.38	3.22
10 x 10	0.015	0.66	4.64
15 x 15	0.023	1.12	12.34
65 x 65	0.028	2.32	20.26
	Power deterrence function		
3 x 3	0.007	0.66	0.98
6 x 6	0.029	2.20	4.88
10 x 10	0.027	2.22	10.82
15 x 15	0.050	4.15	17.59
65 x 65	0.086	7.58	27.19

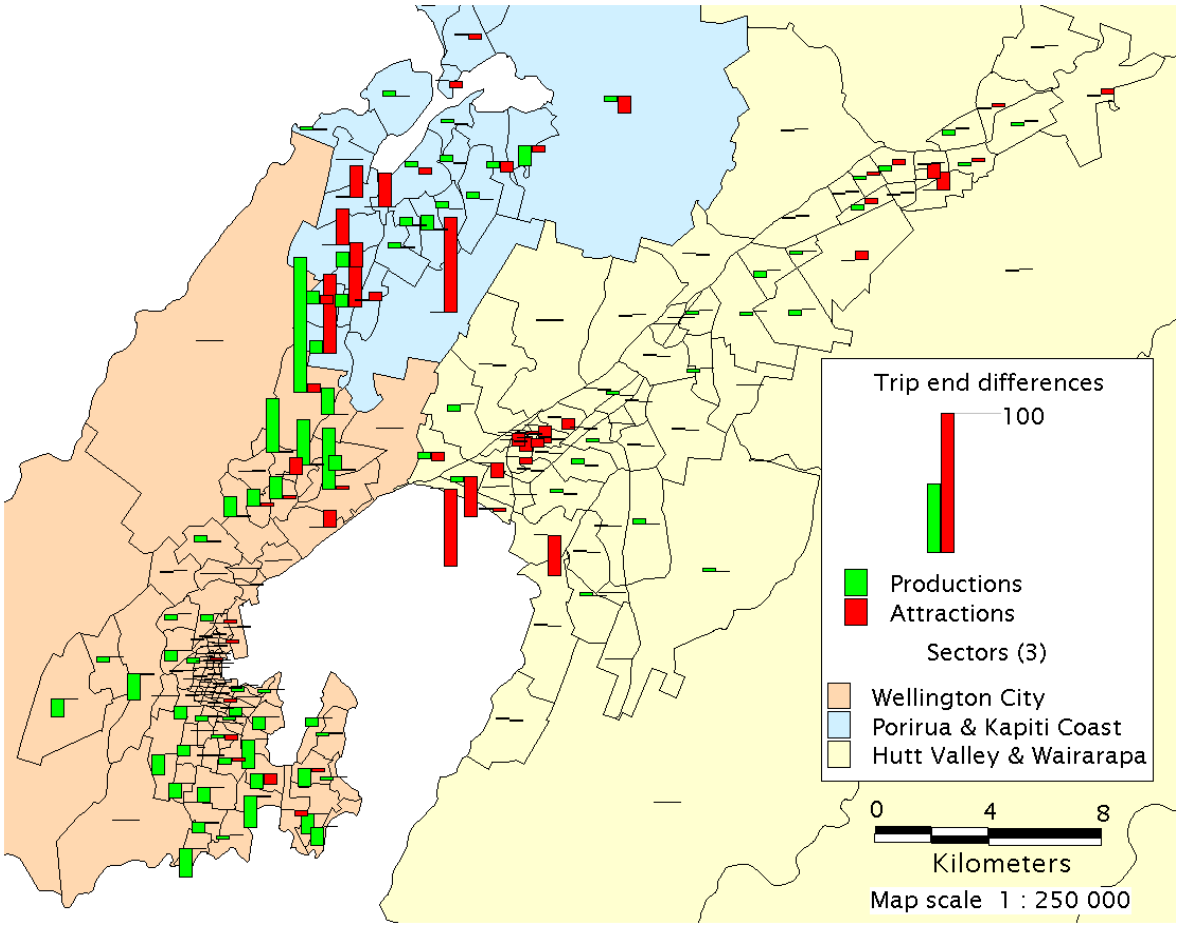
Table 8.18 shows the deviances between the fitted and observed trip end totals summed over both productions and attractions. They include weighting according to their confidences; even so, deviances are smaller for strong confidence and greater for weak. These deviances are added in to the total residual deviances from trip distribution models (eg table 8.14) and detract from the changes in deviance with cost (table 8.15), but are generally small compared with either, particular the changes in deviance.

To compare the mismatches on a common scale, the weighting can be removed by dividing the strong deviances by 100 and multiplying the weak deviances by 100. This increases the relative differences even further.

The mismatch generally increases with increasing disaggregation down the table. This is despite any constraint on the zonal trip end totals by the increasing number of sectors, whose trip end totals are determined by summation of segments. It is possible that the reducing size of sectors limits the range of costs between zones within the sector, and weakens any linkage between trip distribution effects and the zonal trip ends within sectors.

8.7.3.8.1 Spatial pattern of trip end differences

Figure 8.5 Differences between observed and fitted trip ends



Daily HBW person-trips by car, from fitting Exponential model to 3×3 segments with ordinary trip end weights

Trip end mismatches appear to vary quite smoothly within the sectors used for aggregation, with abrupt changes across their boundaries. This is in contrast with convergence problems, where errors are concentrated on a few remote zones.

There is a general balance between positive and negative errors within each sector, but there are still small errors in sector trip end totals.

See also figures 8.9 and 8.10 for trip end mismatches with actual screenline counts.

8.7.3.8.2 Confirmation of best fit solution

The Exponential model estimated from 3x3 segments was checked by perturbing the cost coefficient and by forcing consistency with observed trip ends by a Furness process. The original estimated matrix had a lower residual deviance than any of these adjustments.

The fits of all estimated matrices were calculated at all levels of aggregation. The best fit at a given level of aggregation was always achieved by the matrix estimated at that level. In particular, GLMs calculated on zonal data gave the best fit at that level but not to aggregated data. The measure of fit was the sum of deviances for aggregated segments and for zonal trip ends, in keeping with the objective function of the matrix estimation.

8.7.3.8.3 Theoretical mechanism and practical responses

Theoretical mechanisms by which the pattern of zonal trip ends within sectors can be related to trip distribution effects are discussed in section 8.5.3.

Trip end mismatches at the sector level are small, probably because they are constrained here by the full set of segment totals as well as zonal trip ends. In practice, only some segment totals are observed from screenline crossings and sector trip ends are needed separately for determinacy. In the practical case of calibration from screenline counts in Wellington, mismatches at the sector level are larger though still quite modest. Their implication for the fit of synthetic demand models is discussed in section 8.7.7.6 and possible responses are suggested in section 8.7.7.7.5.

In either case, the mismatch of zonal trip ends within sectors is probably too small to be of practical importance.

8.7.3.9 Indeterminate models

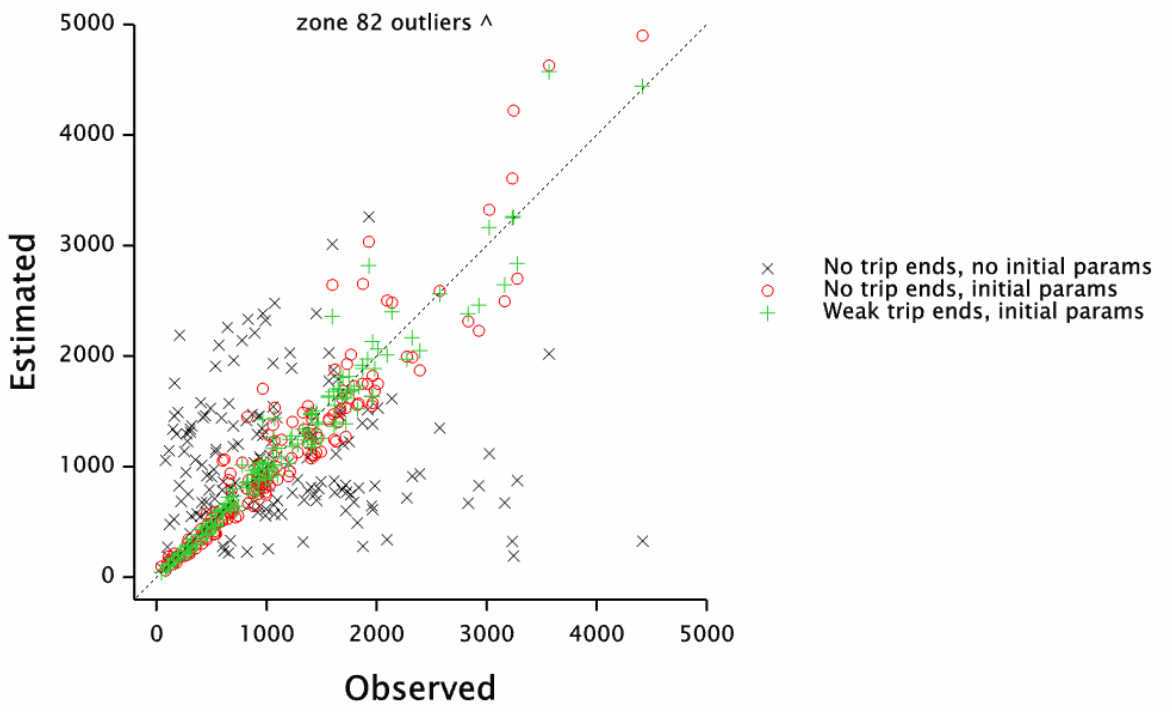
Trip end data was included, with various confidences, to avoid indeterminacy in the allocation of trips between zones within sectors. Without trip end data, models would be indeterminate and this was expected to lead to severe problems in running MVESTM.

It was found that such models would run without catastrophic failure. There were indications of problems such as the fixing of parameters and increases in the number of iterations, but fitted cost coefficients and the changes in deviance associated with them were plausible. In general, these results continued the trends seen with weakening confidence in trip end data, with greater variation in cost coefficients between different levels of aggregation.

One point of interest was that the fit to observed data aggregated to 3x3 segments was almost perfect, with a very small residual deviance for all three deterrence functions. Such a fit is quite plausible given the under-specification of the model. Such solutions should also exist for the other aggregations, apart from 65x65 which is simply determinate, but MVESTM did not reach them. These measures of deviance exclude the fit at trip ends since no trip end data was included in the models.

In these cases the fit to zonal trip ends was remarkably good, given that no trip end data file was input. It was suspected that information was being derived from the initial values provided for trip end parameters. When these were reset to unity, rather than values for a fitted trip distribution (section 8.3.6), the fit to zonal trip ends became worse, as would be expected in the absence of any information about them. The fit of trip ends for an Exponential model calibrated to 3x3 segmentation is shown in figure 8.6.

Figure 8.6 Fit of weak and indeterminate trip ends



The figure shows the fit is much improved by initialising the trip end parameter values, but still not as good as when trip end data is introduced with weak confidence. The fit when trip end data is introduced with ordinary or strong confidence is so good that all points appear to lie on the diagonal of this figure. Outlying points for zone 82 fall above the top of the plot in all three cases; zone 82 covers Churton Park and Glenside north of Johnsonville, on the sector boundary.

In most cases, where an exact fit to segment totals was not found, the trip ends from indeterminate models were more scattered.

8.7.3.10 Complementary within-segment calibration by K factor

These MVESTM models are fitted to trip totals for segments. This type of fitting was considered when developing sector systems for geospatial analysis, but there does not appear to be any way of aggregating costs to segments consistently (section 7.6). MVESTM overcomes this by modelling at the zonal level.

The trip distributions fitted by MVESTM are calibrated on the differences between segments. Information within segments is lost in the aggregation of trip totals.

Trip distributions can be calibrated on this within-segment information, by fitting a GLM to the zonal observed matrix with a separate constant or K factor for each segment. The K factors provide an exact fit to the trip totals for each segment, absorbing between-segment effects and leaving cost coefficients to be fitted to within-segment effects alone.

The K factors are fixed effects. They not expected to conform to a random distribution, as is the case for the random K factors fitted by HGLM in chapter 7.

Table 8.19 shows that for a flat model, without any trip distribution effects of cost, the residual deviances from fitting between- and within-segment are complementary. Outputs from the two software packages show good consistency in numerical precision.

Table 8.19 Between and within segment residual deviances for flat model

Segmentation	Between segment from MVESTM	Within segment GLM with K factors	Total
1 x 1– none	~	6380.441	6380.441
3 x 3	959.467	5420.972	6380.439
6 x 6	1661.957	4718.483	6380.440
10 x 10	1979.705	4400.739	6380.444
15 x 15	2136.370	4244.075	6380.445
65 x 65	3641.286	2739.145	6380.430
225 x 225 – zonal	6380.441	~	6380.441

The deviances for the between-segment model fitted by MVESTM are independent of trip end weighting, because flat models fit trip ends exactly.

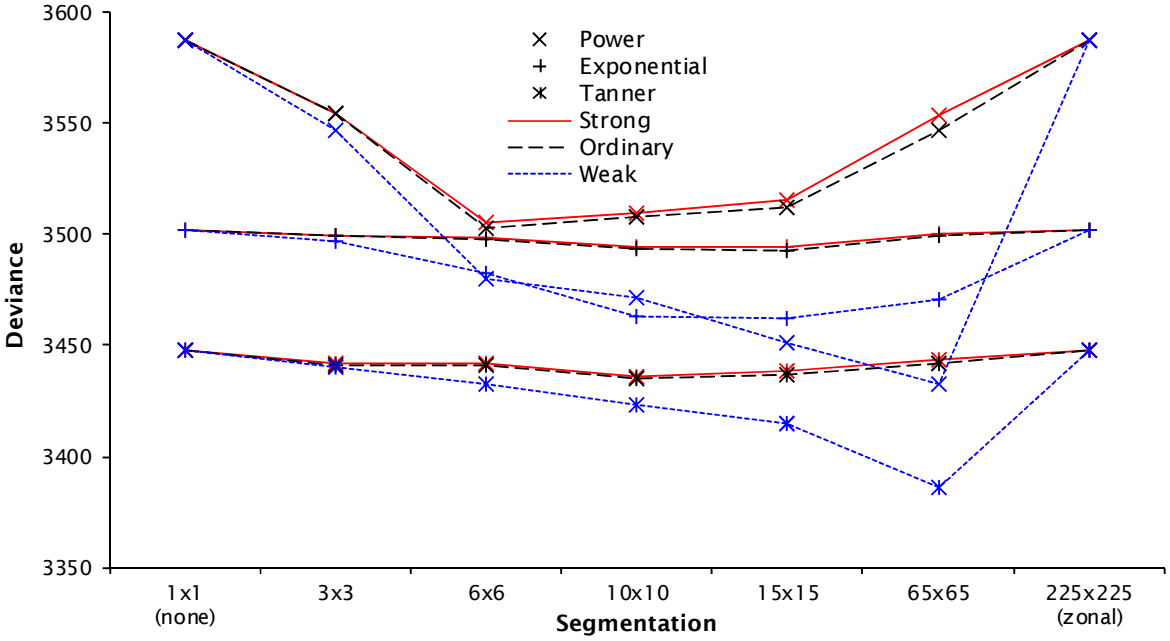
The fitted trip matrices are the same, by either approach, for all segmentations. Differences in the deviances arise from the different data on which they are fitted. This data is the segment totals for MVESTM, and trip patterns within each segment for the GLM once the segment totals are fitted out by the K factors.

Once cost deterrence is introduced, different models are fitted to the between- and within-segment information. The two models are complementary but not consistent; each optimises its fit to one part of the dataset, ignoring the fit to the other part.

Figure 8.7 plots the sum of the residual deviances from the two approaches. This gives the best of both worlds, although it cannot be provided by any single model. The best single model is the simple GLM fitted to the zone-to-zone observed trips without K factors.

At the top of table 8.19 and on the left of figure 8.7, there is just one segment for the whole study area so all calibration is within-segment. At the bottom of the table and to the right of the figure, each segment is a zone-to-zone movement, and all calibration is between these segments. In both cases the calibration is by a simple GLM.

Figure 8.7 Sum of residual deviances from between and within segment calibration



This single GLM provides the endpoints for the plots in figure 8.7. The dip in between represents an improvement in fit by treating within- and between-segment effects separately. It can be seen as the failure of a single model to fit both levels of a spatial hierarchy simultaneously.

Under this interpretation, the Exponential deterrence function allows the most consistent fit to both levels at once. Reductions in deviance are no more than 11.2 for ordinary trip end weighting; while this is significant (against χ^2_1), it is small compared with the reduction when cost deterrence effects are introduced to the flat model, 2130.9.

The Tanner deterrence function provides a less consistent fit to both levels of the hierarchy, but it is the Power function where there is the most marked reduction in summed deviance by fitting separate models to the two levels. Although the plot shows these summed deviances can be less than for the Exponential or even the Tanner function, no single model can achieve this fit. Single Exponential or Tanner models (represented at the endpoints) can generally achieve as good a fit to all the data as inconsistent Power models fitted between- and within-segments separately.

Strong and ordinary trip end weights give very similar results throughout, suggesting that even the ordinary weighting is close to the limiting case of fixed trip ends. The much reduced sum of deviations with weak trip ends may be achieved by changes in trip ends from their observed values with MVESTM. This is most marked at 65x65 segmentation, where inconsistencies are noted in section 8.7.3.6.1, and the Tanner mimics the Power to a lesser extent.

More than two levels of hierarchy could be investigated in this way. Intermediate hierarchies could be fitted by MVESTM to trips totalled at the lower level of segmentation with K factors (introduced as dummy screenlines, w = 0) for the higher level of segmentation.

8.7.4 Calibration on a single screenline or matrix segment

The principle advantage of matrix estimation is its ability to estimate from aggregate data that can be easily observed, eg the count of cars crossing a screenline. A simple but artificial case of this can be developed from the calibration of matrix data described above. The data is already aggregated into segments; in the coarsest case just nine segments between three sectors. Data for all segments, representing the whole matrix, is used in the calibration above. The data is introduced to MVESTM in its terminology as 'screenlines'.

To represent a single screenline, data for just one segment is included in each calibration below. Each of the nine segments of the 3x3 segmentation has been calibrated this way. In practice, it is hard to observe intrasector movements if screenlines follow the sector boundaries.

8.7.4.1 MVESTM formulation

Table 8.20 MVESTM formulation for calibration on a single screenline or segment

Data	Scope	Structure
Prior matrix		
Costs		$t_{ij} = c_{ij}^{-\gamma}\exp(-\lambda c_{ij})$
Trip ends		**Parameters**
Totals from trip matrix to provide proportions within sectors, and differences from 'counted' segment		Free to fit to trip end total x balancing factor
Screenlines	**Intercepts**	
PA trip data for one segment representing a screenline count	One segment	*Fixed at unity*
		Alpha, beta free to calibrate $-\gamma$, λ

Italic – null effect on fitted model
Note: cf table 8.2 or E.1

8.7.4.2 Fit of model

Trip end data is again included. With the trip total for only one segment as additional data, there is just sufficient data for a determinate model with one cost coefficient. The Tanner model cannot be fitted uniquely and the Exponential and Power models have no residual degrees of freedom. The fitted residual deviance was typically 10^{-5} or less.

The notable exception was the synthetic Exponential model from the Kapiti to City segment with strong trip end confidence, with a total residual deviance of 0.15 – still a good fit in practical terms.

The distributions replicate the segment trip totals well, typically to within one trip in many thousands, but again with exception of the Kapiti to City segment which had errors up to 600 (5%).

Different confidences were again applied to the trip end data. Reports from the optimisation process of parameters not contributing to the estimation suggest problems with fitting and convergence. There were no such reports for strong trip end confidences. For ordinary confidences, the production parameter for zone 41 was reported for every distribution model: zone 41 has just one observed trip production. With weak trip ends, there were 600 reports for various trip ends, mainly productions, among the models.

8.7.4.3 Information about trip distribution

Since the distribution models are just determinate with negligible residual deviance, the change in deviance, which reflects the information about distribution effects, is equal to the deviance of the null models. These differ for each segment because the inclusion of that segment's trip total forms a different dataset. However, there is no distinction between Exponential and Power deterrence functions; these are absent in the null model and remove all the mismatch in the fitted model, irrespective of form.

Table 8.21 Change in deviance for single HIS segments

Segment		Trips	Ratio to flat x	Trip ends			
Production (home)	Attraction (work)			Fixed	Confidences		
					Strong	Ordinary	Weak
City	City	57,890	1.92	126.2	125.3	73.3	1.7
City	SH1 Kapiti	3218	0.33	37.5	37.4	28.6	1.3
City	SH2 Hutt	6133	0.22	152.0	151.4	111.6	4.7
SH1 Kapiti	City	12,552	0.71	10.7	10.7	7.3	0.2
SH1 Kapiti	SH1 Kapiti	22,299	3.90	174.4	173.4	111.9	2.9
SH1 Kapiti	SH2 Hutt	4622	0.29	71.6	71.3	53.4	2.3
SH2 Hutt	City	11,882	0.35	125.0	124.5	87.6	3.0
SH2 Hutt	SH1 Kapiti	1018	0.09	96.7	96.3	76.2	4.5
SH2 Hutt	SH2 Hutt	63,602	2.05	165.4	164.2	94.9	2.2
Total		**183,216**	**1**	**959.5**	**954.6**	**644.9**	**22.8**

Strong trip ends favour the common flat model that is completely consistent with the trip ends alone. Most of the deviance arises from the departure of the segment trip total from this model.

As the trip ends weaken, the individual flat models adjust to accommodate the segment trip total at the expense of the trip end data. However, this is a relatively small expense due to the weakness of the trip end data, and leaves little to be improved by the introduction of a cost component representing distribution effects.

The marked reduction in change of deviance as trip ends weaken follows the pattern plotted for a minimal case in figure 8.3.

The limiting case of stronger trip ends is fixing the trip ends. This can be taken from the flat model including all the segments, fitted in table 8.14, top line, left column. In combination, the segments are consistent with the trip ends, so the common flat model is fitted exactly to the trip ends. All deviances arise from the lack of fit of individual segments to the common flat model. These are shown for the individual segments in the left-hand data column of table 8.17. They are slightly greater than the deviances of models fitted to individual segments and strong trip ends, because these allow some adjustment to trip ends.

Fitting to synthetic datasets shows similar patterns.

8.7.4.4 Cost coefficients

Cost coefficients recovered from synthetic datasets are typically within about one part per thousand of the original and even better with stronger trip ends. The Kapiti to City segment is again an exception, with some coefficients returned close to their initial values.

The models fitted to observations are just determinate, so their coefficients should be the same for irrespective of trip end weights. The coefficients for strong and ordinary trip ends are close, to within one part per thousand, and are shown in table 8.22. With weak trip ends, there are differences of up to 3% or more for the Kapiti to City.

Table 8.22 Cost coefficients fitted to single HIS segments

Segment		Fitted cost coefficient	
Production (home)	**Attraction (work)**	**Exponential λ**	**Power γ**
City	City	0.065	1.58
City	SH1 Kapiti	0.054	1.37
City	SH2 Hutt	0.070	1.69
SH1 Kapiti	City	0.050 (0.048) *	1.23
SH1 Kapiti	SH1 Kapiti	0.063	1.64
SH1 Kapiti	SH2 Hutt	0.072	2.04
SH2 Hutt	City	0.069	1.74
SH2 Hutt	SH1 Kapiti	0.076	2.00
SH2 Hutt	SH2 Hutt	0.071	1.81
Trip weighted average		**0.066**	**1.67**

* Poor convergence suspected. Correct value may be around 0.048. See section 8.7.4.5.

Cost coefficients fitted to observed data vary between segments, indicating that the segment totals do not follow a common model form. Power coefficients vary relatively more between segments than those for the Exponential model. The average, weighted by observed trips, is reasonably close to the coefficient fitted to all segments in a single model (3x3 aggregations in table 8.17).

8.7.4.5 Convergence with the Kapiti to City segment

There are several signs of poor convergence of models based on the Kapiti to City segment. When synthetic datasets were calibrated starting from alternative initial values of the cost parameter, the fitted values were often closer to the initial values than to the known value used to synthesise the dataset. This tended to occur more for Exponential deterrences and weak trip ends. Examining Exponential coefficients fitted to observed data from varying initial values suggested a true value around 0.048.

In two cases with initial values relatively distant from their expected values, and with strong trip ends, cost parameters were reported as not contributing to the estimation.

This is not a simple matter of small sample size, since the segment is the largest of the six intersector movements. However, the deviances for the segment shown in table 8.21 are much lower than for other segments; this is also the case for synthetic data. The ratio of the segment trip total to the expectation from a flat model based on trip ends alone, x=0.71, is much closer to unity than for any other segment. Figure 8.2 shows how this can lead to the low deviances and it seems likely that this lack of contrast is also responsible for poor convergence.

8.7.5 Quadrants and diagonal pairings

The previous analyses show that trip distribution models can be calibrated from the trip total for a single segment, which can be seen as a form of screenline observation. In practice, a single PA matrix segment

may not be distinguishable in traffic counts but a two-way traffic count can give the total for two segments, one transposing the production and attraction sectors of the other.

To examine the consequences, the HIS matrix is aggregated further into just two sectors, giving four segments or quadrants. This is the minimal case for trip distribution effects, with the two sectors separated by a single screenline at which movements between them can be observed. It is the basis for the theoretical considerations in section 8.5.

The example combines the SH1 Kapiti and SH2 Hutt sectors of the 3x3 aggregation into a single sector, Kapiti & Hutt, leaving the city of Wellington, including Johnsonville, as the other sector.

Since all data is drawn from the same HIS matrix, the trip total for one quadrant determines those for all others, given the sector trip ends. With this internal consistency, all formulations are fitted exactly by the same trip distribution models, either Exponential $\lambda = 0.065$ or Power $\gamma = 1.58$. These are the models fitted to the City to City segment, common to both 2x2 and 3x3 aggregations, in the top line of table 8.22. Residual deviances are less than 0.01, often much less.

The first four rows of table 8.23 show the information obtained from fitting each the four quadrants separately. The first row, for the intrasector city movements, corresponds with the first line in table 8.21. The fifth line totals the change in deviance from all four calibrations.

Pairs of diagonally opposite quadrants are then fitted together. The first pair is the leading diagonal, comprising the intrasector movements. The second pair on the trailing diagonal comprises the intersector movements, which can be observed crossing a screenline between the sectors.

For each pair of diagonally opposite quadrants, changes in deviances are shown from:

1 the sum of two separate calibrations, from the top four lines

2 a single calibration with the two quadrant trip totals entered as two separate items

3 a single calibration with the two quadrant trip totals aggregated into a single item of information.

Table 8.23 Change in deviance for HIS quadrants

Quadrant(s)		Trips (HIS)	Proportion to flat x	Deviance			
				Trip ends			
				Fixed	Strong	Ordinary	Weak
City to City		57890	1.916	126.2	125.3	73.3	1.7
City to Kapiti & Hutt		9351	0.253	187.6	186.7	132.3	4.7
Kapiti & Hutt to City		24434	0.469	116.1	115.5	76.3	2.2
Kapiti & Hutt to Kapiti & Hutt		91541	1.433	66.9	66.4	38.3	0.8
Total		183216	*	496.8	494.0	320.2	9.5
City to City and K&H to K&H	Sum	149431	*	193.2	191.7	111.6	2.5
	Separate		*	193.2	192.6	147.2	6.2
	Aggregate		1.588	174.7	174.1	128.5	4.2
City to K&H and K&H to City	Sum	33785	*	303.7	302.3	208.6	6.9
	Separate		*	303.7	303.1	251.5	11.9
	Aggregate		0.379	286.0	285.4	234.7	10.7

K&H - Kapiti & Hutt

The total of trips in each quadrant or combination is shown, together with its ratio to the expectation from the flat model determined by trip ends. This is the parameter x discussed in section 8.5.4. The next column shows the deviances of the quadrant totals from this fixed initial model. The final three columns show the deviances with strong, ordinary or weak trip end confidences.

As with individual 3x3 segments, there is little loss of deviance between the fixed and strong trip ends. There is a more substantial reduction to ordinary trip ends, but the major part of the deviances are retained and are still hugely significant. By contrast, they are barely significant with weak trip ends.

With fixed trip ends, the deviances of individual quadrants are simply additive in combination, except where they are aggregated and there is a loss of deviance. As trip ends weaken and deviances reduce, the inclusion of two quadrants separately provides most information. For ordinary and weak trip ends, the deviance from one aggregate analysis is greater than the sum from analyses of the individual quadrants. This symbiotic effect may be due to the diagonal pair of quadrants providing a scale for the whole matrix which is being lost with weakening trip ends.

For weak trip ends, the deviance from the observable quadrants in combination, even aggregated (10.7), is greater than the total from all quadrants singly (9.5).

Loss in aggregation is only a minor part of the deviance available from diagonal quadrants.

In general, the deviance is more closely related to the strength of the contrast x (ie its difference from unity) than to the number of trips; this accords with the influence of x in special minimal cases shown in figure 8.2. In particular, deviances from the observable intersector quadrants are larger than from the intrasector quadrants, although there are far more trips in the latter.

There is a marked loss in the total deviance change for this 2x2 aggregation compared with 3x3 aggregations in table 8.15, or the bottom line of table 8.21. It is about the halving that might be expected simply from using data at one screenline between two sectors, rather than two screenlines between three sectors.

8.7.6 Transition to an actual screenline count

The movements between the City and the Kapiti & Hutt sectors, analysed in section 8.7.5, are similar to those intercepted by the radial screenline. This allows a comparison between the previous somewhat artificial calibrations on household travel survey (HIS) trips aggregated by the segments and more realistic calibrations from traffic counts on the screenline. Table 8.24 follows differences between the two approaches step by step.

Table 8.24 Changes from quadrants of HIS data to screenline counts

Change to:	Screenline		Deviance	Cost coefficient	
	Count	Confidence		Exponential λ	Power γ
Segments (base)	33,785	0.633	234.7	0.0650	1.577
Screenlines	38,816	0.633	270.5	0.0689	1.640
Hourly flows (3)	11,006	2.23	255.7	0.0701	1.649
Confidences	11,006	6.07	518.7	0.0701	1.649
– trip end data	11,006	6.07	590.7	0.0707	1.613
– screen data	10,783	6.07	535.6	0.0720	1.682
All data	10,783	6.07	608.8	0.0725	1.644

Calibrations are made on the single ‘screenline count’ shown in the first column, representing an aggregation of all periods and directions. The corresponding confidence is in the second column. With

only one item of information about the trip distribution, only a single parameter can be fitted for the deterrence function. This achieves an exact fit, so the deviance is that of the flat model and is interpreted as the amount of information about trip distribution.

Each successive line of the table introduces a change in one aspect of the differences between the HIS dataset and the screenline count

Segments

This base case corresponds with the last row in table 8.23 for ordinary trip end weighting. The count is an aggregation of all intersector movements (ie the trailing diagonal quadrants) between the City and the Kapiti & Hutt sectors, taken from the 24-hour PA matrix observed from the HIS.

Screenlines

This is a geographical change to the radial screenline. The movements are now defined by the intercept files from select-link analyses of the screenline. These differ by period, mainly because of alternative routings between Porirua and the Hutt Valley, and in this case the intercept proportions have been simply averaged across the three periods. Directional intercepts have similarly been averaged and transposed to abstract a 24-hour count from the HIS observed matrix.

Hourly flows

This is a change in time period formulation. Six one-hour flows representing screenline crossings by period and direction are abstracted from the 24-hour PA HIS matrix, using the same WTSM period and direction factors incorporated into MVESTM intercept file (see section 8.6.3.3 and appendix C.2). The six flows are totalled into a single count.

Because this represents an averaging by period of the all-day sample of trips from the HIS, the confidence is adjusted to give the same product with the count as for the previous row, ie

$$38816 \times 0.633 = 11006 \times 2.23$$

This approximation maintains the derivation of the confidence from the effective HIS sampling rate.

Confidence

This is a change in confidence to that expected from screen line counts (section 8.4.4.2)

For simplicity and want of better information, the same ordinary trip end confidences are used throughout, 100/157.9 = 0.633. This is based on the effective sampling of the HIS and adopted for synthetic trip ends as lying within the likely range.

Trip end data

This changes the trip ends from data observed in the HIS to data synthesised in the WTSM model.

Screenline data

This changes the source of screenline data from aggregations of trips in the HIS to traffic counts on the screenline, but retains the trip ends observed in the HIS.

All data

This combines traffic counts at screenlines with synthetic trip ends. No observations are taken directly from the HIS, so this represents a calibration of trip distribution without recourse to such expensive disaggregate data.

The deviance for this final combination is substantially higher than from the HIS. The difference arises mainly from the higher confidence in the screenline counts than in the HIS. However, this change in

deviance is not directly proportional to the change in screenline confidence, since the calibration depends on contrasts with trip end information.

The large final deviance suggests considerable power in screenline counts for modelling trip distribution, and a good margin for error in confidences. It compares well with the reduction in deviance of 2130.9 when fitting an Exponential model to the fully disaggregate household data.

Some increase in the deviance also occurs with the change from segments to the screenline, perhaps simply because more trips are intercepted.

The main difference in cost coefficients appears with the geographical change from segments to screenlines. The coefficients in the table are generally closer to each other than to the coefficients fitted to the fully disaggregate household matrix, λ=0.0638 and γ=1.398; this could be an affect of aggregation, as seen in table 8.17, as much as a local variation in the trip distribution.

There is no change in the coefficients when the confidence in the screenline changes.

The effects of changes in screenline and trip end data appear additive in both deviance and cost coefficients.

8.7.7 Calibration on actual screenline counts

The transition from aggregations of HIS travel data to actual screenline counts in the previous section was on the convenient common ground of an all-day, two-way count at one screenline (radial). This section considers calibrations on four screenlines, by three periods and in two directions. It addresses practical issues of period and direction posed by real screenline counts.

8.7.7.1 Formulation

Table 8.25 MVESTM formulation for calibration on actual screenline counts

Data	Scope	Structure
Prior matrix		
Costs	24-hour PA	$t_{ij} = c_{ij}^{-\gamma}\exp(-\lambda c_{ij})$
Trip ends		**Parameters**
Synthetic generations 24-hour PA		Free to fit to trip end total x balancing factor
Screenlines	**Intercepts**	
Observed hourly directional counts, by period (factored down to internal HBW using model)	Real screenlines, R_{ijk} from select link analysis, factored for period, direction and occupancy	*Fixed at unity*
		Alpha, beta free to calibrate $-\gamma$, λ

Italic - null effect on fitted model
Note: cf table 8.2 or E.1

The trip distribution is estimated as a 24-hour PA person trip matrix, as previously. Synthesised trip ends are provided on the same scale; they are abstracted from the WTSM base-year model after the joint distribution and mode split to give car trip ends.

Counts are presented as hourly vehicle flows, independent between screenlines, periods and directions. Period, direction and occupancy factors relating them to the 24-hour production–attraction person trip matrix are incorporated in the intercept proportions (see section 8.6.3.3).

8.7.7.1.1 Individual, separate and aggregate counts

Results from calibrations on each individual combination of screenline, period and direction are shown in the upper left body of tables 8.26 and 8.27. The margins of the tables, to the right and below, give results including all screenlines, periods or directions. These are entered into the analysis either as separate items of information or aggregated into a single total. The same data is presented in either case but in a different form.

8.7.7.2 Fit of model

Where counts are presented singly or as a single aggregation, fitting a single cost coefficient can achieve an exact fit. The residual deviance in these cases is small, typically of the order of 10^{-9}. There are a few cases of less exact convergence to around 10^{-2}, mainly for the small IP counts on the Rimutaka screenline. A few Exponential models aggregated across screenlines did not fit any attractions to zone 224 in the Wairarapa where there were 136 synthetic trip ends, giving an excess deviance of 20. This was resolved by a better initial value for the zone's balancing factor in the parameter file.

An exact fit cannot be expected with a single cost coefficient when several counts are presented separately and the residual deviance is a measure of model fit. The mean residual deviance for Exponential models is generally less than unity, suggesting that higher confidences may be justified and that a synthetic trip distribution is able to meet EEM and DMRB model validation criteria.

For Power models, mean residual deviances tend to be greater than unity where screenline counts are entered separately, but not where only directions and period counts are entered separately. This suggests that trip distribution effects appear as contrasts between screenlines rather than between periods and directions.

Total residual deviances for Exponential models are generally too low to allow more complex models such as the Tanner function or geographic segmentation to demonstrate significance in an improved fit.

8.7.7.3 Information about trip distribution

Even where there are residual deviances, they are small in comparison with the reduction in deviance when cost deterrence is introduced, so the Exponential and Power deterrence functions show broadly the same pattern. The changes for the Exponential are shown in table 8.26.

Table 8.26 Change in deviance with fitting Exponential trip distribution to screenline counts

Direction	Period	Screenlines				Aggregate	Separate
		Central	Radial	Regional	Rimutaka		
Inbound	AM	84	163	44	112	266	326
Inbound	IP	19	34	17	35	84	102
Inbound	PM	150	144	69	114	379	414
Inbound	Aggregate	172	349	103	232	563	706
Inbound	Separate	212	362	116	240	643	783
Outbound	AM	201	187	93	160	478	533
Outbound	IP	19	40	17	33	90	106
Outbound	PM	80	133	29	90	240	280
Outbound	Aggregate	226	385	114	255	692	814
Outbound	Separate	257	385	129	262	728	858
Aggregate	AM	205	403	121	262	660	815
Aggregate	IP	38	73	32	64	163	195

Direction	Period	Screenlines				Aggregate	Separate
		Central	Radial	Regional	Rimutaka		
Aggregate	PM	180	299	84	195	547	642
Aggregate	Aggregate	273	609	157	363	849	1091
Aggregate	Separate	274	609	157	363	849	1091
Separate	AM	254	410	138	274	739	896
Separate	IP	38	73	32	64	163	195
Separate	PM	205	301	95	200	591	684
Separate	Aggregate	279	610	157	364	862	1099
Separate	Separate	337	620	181	379	953	1200

All changes in deviance are highly significant; they are in double figures even for the small directional IP counts at individual screenlines. The changes in deviance are not simply proportional to the traffic counts. In many cases, deviance changes for counterpeak movements (AM out and PM in) are greater than those for the corresponding peak movements (AM in and PM out), contrary to the pattern of counts.

The overall trip distribution information represented by the change of deviance of 1200 is more than half that available from the HIS at the zonal level, in which there is a change of deviance of 2131 (Exponential) or 2197 (Tanner).

8.7.7.3.1 Aggregation and separation in period and direction

If directions and periods have been aggregated, effectively presenting a single all-day two-way count, no more information is gained by separating periods and hardly any by separating directions. There is a relatively modest increase in information if counts are separated by both period and direction, more so at the central cordon.

This is because, broadly speaking, the same set of PA movements will be intercepted in each direction over 24 hours, so there is little contrast between them when separate directional counts are presented, and little gain in trip distribution information. Similarly, the same set of PA movements is intercepted by two-way counts in both the morning and the evening peak. The contrast between peak and counterpeak movements only appears when both directions and periods are separated.

The amount of information thus gained is relatively modest, suggesting that the calibration may be relatively insensitive to the period and direction factors (eg table 8.7) incorporated in the intercept proportions. It also suggests that relatively little information may be lost by working with two-way counts in an all-day model, which may be simpler and more consistent with the rest of a demand model.

8.7.7.3.2 Aggregation and separation over screenlines

Deviance changes for combinations of counts tend to be less than the sum of the individual counts, but there is a reasonable gain in presenting the screenlines separately rather than in aggregate. The combination of IP counts comes closer to the sum of their parts possibly because the small counts are the major limit on information rather than the accuracy of trip ends.

8.7.7.4 Cost coefficients

Table 8.27 Fitted cost coefficients – screenline counts

Direction	Period	Screenlines				Aggregate	Separate
		Central	Radial	Regional	Rimutaka		
		Exponential deterrence function					
Inbound	AM	0.068	0.089	0.050	0.148	0.074	0.079
Inbound	IP	0.068	0.070	0.051	0.236	0.069	0.070
Inbound	PM	0.069	0.061	0.055	0.074	0.064	0.063
Inbound	Aggregate	0.068	0.075	0.049	0.119	0.070	0.073
Inbound	Separate	0.069	0.067	0.053	0.101	0.066	0.067
Outbound	AM	0.068	0.062	0.056	0.102	0.065	0.064
Outbound	IP	0.064	0.069	0.051	0.236	0.067	0.067
Outbound	PM	0.071	0.087	0.040	0.117	0.074	0.081
Outbound	Aggregate	0.069	0.070	0.048	0.118	0.068	0.071
Outbound	Sep	0.068	0.068	0.052	0.115	0.067	0.067
Aggregate	AM	0.068	0.073	0.050	0.123	0.069	0.072
Aggregate	IP	0.066	0.069	0.051	0.236	0.068	0.069
Aggregate	PM	0.070	0.072	0.045	0.092	0.069	0.073
Aggregate	Aggregate	0.069	0.072	0.048	0.119	0.069	0.073
Aggregate	Separate	0.069	0.072	0.048	0.116	0.069	0.073
Separate	AM	0.068	0.068	0.054	0.116	0.067	0.067
Separate	IP	0.066	0.069	0.051	0.236	0.067	0.069
Separate	PM	0.069	0.067	0.051	0.088	0.066	0.067
Separate	Aggregate	0.069	0.071	0.048	0.118	0.068	0.071
Separate	Separate	0.068	0.068	0.053	0.110	0.067	0.067
		Power deterrence function					
Inbound	AM	1.428	1.921	1.738	3.926	1.683	1.765
Inbound	IP	1.374	1.618	1.891	3.967	1.604	1.759
Inbound	PM	1.433	1.449	1.982	2.790	1.537	1.577
Inbound	Aggregate	1.426	1.683	1.826	3.657	1.620	1.828
Inbound	Separate	1.430	1.559	1.940	3.339	1.569	1.667
Outbound	AM	1.420	1.460	2.028	3.244	1.545	1.588
Outbound	IP	1.306	1.588	1.810	3.925	1.545	1.642
Outbound	PM	1.479	1.901	1.555	3.921	1.683	1.767
Outbound	Aggregate	1.434	1.606	1.787	3.553	1.594	1.744
Outbound	Sep	1.423	1.564	1.923	3.553	1.573	1.659
Aggregate	AM	1.424	1.653	1.860	3.709	1.611	1.824
Aggregate	IP	1.341	1.603	1.849	3.968	1.575	1.711
Aggregate	PM	1.457	1.639	1.724	3.251	1.607	1.747
Aggregate	Aggregate	1.430	1.644	1.806	3.604	1.607	1.876

Direction	Period	Screenlines				Aggregate	Separate
		Central	Radial	Regional	Rimutaka		
Aggregate	Separate	1.432	1.642	1.812	3.755	1.606	1.876
Separate	AM	1.422	1.569	1.970	3.607	1.576	1.673
Separate	IP	1.336	1.599	1.841	3.966	1.569	1.696
Separate	PM	1.442	1.557	1.897	3.160	1.569	1.644
Separate	Aggregate	1.431	1.622	1.789	3.576	1.597	1.802
Separate	Separate	1.425	1.562	1.942	3.493	1.572	1.695

Overall, coefficients are larger than those calibrated from the HIS at the zonal level, 0.0638 and 1.398, but are quite similar to some calibrated on aggregations of that data.

The main differences appear between screenlines. The sense of difference can differ by deterrence function; the Exponential coefficients for the regional screenline are low but the Power coefficients are high, whereas for the Rimutaka screenline both Exponential and Power coefficients are high.

Coefficients for the Rimutaka screenline differ quite markedly from others. The count was conducted in the head of the Hutt Valley, rather than on the Rimutaka pass itself. It will have intercepted some shorter distance trips within the Hutt Valley as well as the longer trips between the Wairarapa and the rest of the study area. The count is the opposite side of Akatarawa Road from the RSI sites and may have split zone 135, Maoribank and Timberlea.

Power coefficients are generally higher for separate screenline counts than for aggregations. This effect is less marked for Exponential coefficients, which generally appear less sensitive.

8.7.7.4.1 Tanner function

Individual or aggregated counts do not have sufficient degrees of freedom to fit the two-parameter Tanner deterrence function. Residual deviances from fitting the one-parameter Exponential deterrence function were small, offering little prospect of significance in further fitting.

The most likely prospect was the formulation with all factors presented separately. Its residual deviance of 10.36 from an Exponential model reduced to 7.49 from a Tanner model. The difference of 2.89 is significant at the 10% level for χ^2_1, which tests against the scale of error determined by the confidences. Testing against the empirical residual of 6.17 with 22 degrees of freedom shows high significance. However, this draws on differences between periods and directions which do not appear to contribute much to trip distribution information. A more conservative test against the residual with two degrees of freedom remaining between screenlines shows no significance.

The fitted coefficients are $\lambda = 0.041$ and $\gamma = 0.59$, which are quite similar to those calibrated on the zonal household data, $\lambda = 0.036$ and $\gamma = 0.65$. This shows the usual trading-off between the correlated coefficients, but in the opposite direction to that seen in aggregations of the household data where λ tended to decrease and γ to increase (table 8.17).

8.7.7.5 Sensitivity to trip end confidences

The importance of trip ends was tested by varying their confidences. As before, the confidences were multiplied by 100 for strong trip ends and divided by 100 for weak ones. This is more than the likely range of error in synthesising trip ends and approximates to limiting cases of trip ends as either absolute constraints, or negligible contributors to trip distribution information.

In such extreme cases, convergence was not so good with residual deviances up to 1 where a perfect fit was possible. Blatantly incorrect solutions were given for some Power models with strong trip ends and

since other Power models continued to show a worse fit than Exponential ones, the following table and interpretation are based on Exponential models.

Table 8.28 Sensitivity of deviance changes to trip end confidence

Screenline	Period and direction aggregated				Period and direction separated			
	df	Strong	Ordinary	Weak	df	Strong	Ordinary	Weak
Central	1	703	273	3.5	6	803	337	8.1
Radial	1	987	609	15.2	6	987	620	15.4
Regional	1	373	157	2.4	6	403	181	3.5
Rimutaka	1	764	363	7.0	6	778	379	8.1
Total		2825	1400	28.0		2970	1518	35.1
Aggregate	1	2479	849	12.0	6	2640	953	16.1
Separate	4	2791	1091	21.0	24	2943	1200	25.5

df are degrees of freedom of input count data; change in df is always 1.

Table 8.29 Sensitivity of Exponential cost coefficient to trip end confidence

Screenline	Period and direction aggregated			Period and direction separated		
	Strong	Ordinary	Weak	Strong	Ordinary	Weak
Central	0.0685	0.0685	0.0614	0.0683	0.0682	0.0697
Radial	0.0725	0.0725	0.0649	0.0679	0.0676	0.0647
Regional	0.0482	0.0482	0.0469	0.0518	0.0530	0.0520
Rimutaka	0.1187	0.1187	0.0510	0.0578	0.1097	0.0506
Aggregate	0.0690	0.0690	0.0644	0.0668	0.0666	0.0678
Separate	0.0697	0.0726	0.0663	0.0665	0.0667	0.0663

Where all data is presented for an individual case or in aggregate, only one item of screenline data is entered; this is show by 1df in table 8.28. The fitted trip distribution model is then just determinate and should take the same form irrespective of data weighting or confidences. Table 8.29 shows that the cost coefficients for strong and ordinary trip ends are the same in such cases but those for weak trip ends differ, particularly for the Rimutaka screenline. These are probably a due to incomplete convergence.

8.7.7.5.1 Weak trip ends

These just-determinate cases must rely on trip end data to fit the trip distribution model. The reduction in deviance, typically in the range 3–15, is still an order of magnitude greater than the general limits of convergence, showing that even the weak trip ends are still providing contrasts with single counts that can be significant. Where counts from all screenlines are presented separately, changes in deviance are about 25. This suggests that if there is any trip distribution information in the contrasts between screenlines alone, it is very much smaller than the information available from contrasts with trip ends with ordinary confidences.

8.7.7.5.2 Strong trip ends

Changes in deviance with strong trip ends are about twice those with ordinary trip end confidences. This is a greater difference than seen with aggregated household data in tables 8.21 and 8.23. It is consistent with stronger screenline count confidences making the trip end relatively weaker and shifting the operating region to the right in table 8.3, leaving greater gains to be made from strengthened trip end data.

The change of deviance for all screenlines entered separately is again almost equal to the sum of changes for individual screenlines, as seen in table 8.21 and explained in section 8.5.4.1.

8.7.7.5.3 Separate period and direction

There is relatively little extra information available from period and direction contrasts when their data is entered separately, on the right-hand side of the tables. Trip end confidences make little difference to cost coefficients, except for the Rimutaka screenline where convergence must again be suspect.

8.7.7.6 Fit at screenlines

Table 8.30 shows the fit of models calibrated on all screenlines, periods and directions, all presented separately. Different trip end confidences are shown for an Exponential deterrence function, and a Tanner deterrence function is shown with ordinary confidences. The fit to each of these counts is shown as the GEH.

Table 8.30 Fit at screenlines – GEH

				Exponential						Tanner	
			Observed	Strong		Ordinary		Weak		Ordinary	
Scn	Dir	Per	Volume	Vol	GEH	Vol	GEH	Vol	GEH	Vol	GEH
Cen	In	AM	9730	9789	0.60	9608	1.24	9708	0.22	9548.2	1.85
Cen	In	IP	1076	1083	0.20	1064	0.36	1069	0.23	1048.1	0.86
Cen	In	PM	1910	1973	1.44	1964	1.24	1924	0.33	1907.7	0.05
Cen	Out	AM	2500	2558	1.14	2555	1.09	2498	0.05	2467.9	0.65
Cen	Out	IP	827	815	0.40	809	0.62	811	0.58	790.3	1.29
Cen	Out	PM	6069	6187	1.50	6061	0.10	6105	0.46	6027.4	0.54
Rad	In	AM	3891	4542	***10.03***	4084	*3.05*	3899	0.13	4085.4	*3.07*
Rad	In	IP	480	497	0.75	463	0.79	455	1.15	467	0.61
Rad	In	PM	1474	1327	*3.93*	1424	1.31	1442	0.84	1435.2	1.02
Rad	Out	AM	2035	1862	*3.93*	2032	0.08	2061	0.57	2044	0.19
Rad	Out	IP	350	362	0.63	360	0.54	363	0.66	363.8	0.72
Rad	Out	PM	2551	2965	***7.88***	2677	2.45	2566	0.29	2683.5	2.58
Reg	In	AM	697	544	***6.12***	701	0.17	738	1.54	704	0.28
Reg	In	IP	73	57	1.98	73	0.08	81	0.83	75.3	0.22
Reg	In	PM	100	57	***4.92***	64	*3.99*	96	0.38	83.6	1.71
Reg	Out	AM	119	64	***5.79***	68	***5.33***	118	0.07	99.3	1.89
Reg	Out	IP	61	46	2.05	59	0.23	66	0.65	59.3	0.16
Reg	Out	PM	519	343	***8.48***	440	*3.58*	465	2.41	444.8	*3.37*
Rim	In	AM	401	461	2.88	419	0.88	399	0.11	431.4	1.48
Rim	In	IP	37	55	2.65	50	1.87	46	1.38	51.1	2.09
Rim	In	PM	72	80	0.91	69	0.43	59	1.58	73.8	0.20
Rim	Out	AM	68	107	***4.11***	90	2.43	76	0.88	97.9	*3.25*
Rim	Out	IP	37	54	2.57	49	1.86	46	1.37	49.2	1.89
Rim	Out	PM	255	285	1.84	259	0.22	246	0.57	267.6	0.77
No. of GEH > 4					7		1		0		0
Maximum GEH					10.03		5.33		2.41		3.37
Root mean square GEH					4.18		1.98		0.92		1.63
Cost coefficient					0.0665		0.0668		0.0663		~

Bold – GEH>4
Italics – GEH>3

The square of the GEH corresponds closely with the deviance scaled by the confidence. This is to be expected given the similarity of GEH^2 with Pearson's χ^2 statistic, which is similar to the deviance for numbers larger than unity. After scaling the hourly volumes by the confidence of 6.067%, most have an effective sample size of five or more, although the size is just over two for some volumes across the Rimutaka. This is still sufficient to avoid the more marked effects of sparsity.

Fitting the model to the validation data itself will tend to exaggerate the goodness of fit. The root mean square (RMS) of the GEH can be adjusted from the 24 items of data by one degree of freedom for the cost coefficient fitted to them, giving a divisor of 23, and an increase in the root mean square GEH of 2%. A more conservative allowance would be only two independent items of data from the six period-by-direction counts at each screenline (peak and counterpeak volumes), giving a 7% (√[8/7]) increase in the RMS of the GEH. This is still less than the difference between the Exponential and Tanner functions. Under conventional matrix estimation, there would one parameter fitted for each screenline count, leaving no usable validation information after such adjustment.

With ordinary trip end confidences, only one of the 24 volumes estimated with the Exponential model has a GEH greater than 4, which is the validation criterion for 'most' (EEM) or 'all, or almost all' (DMRB) screenlines. The model is thus at least close to meeting these criteria. The Tanner model, whose improvement over the Exponential appears marginally significant (section 8.7.7.4.1), meets the criteria with a comfortable margin.

This is contrary to the perceptions and findings that pure synthetic trip distributions fit poorly; that they require K (and L) factors, and empirical observed matrices are preferred as base matrices. A good fit has only been demonstrated here for commuting trips by car, with uncertainty as to confidence levels, in the single instance of the WTSM model for Wellington. Satisfactory validation would have to be demonstrated in other models of other study areas to establish any generality.

More general findings may be drawn from the relative effects of strong and weak trip end confidences. With strong trip ends, GEH values are roughly doubled, and fall well outside the validation criteria; the average (RMS) is greater than 4. With weak trip ends, GEH values are roughly halved and meet the validation criteria by a clear margin.

Strong confidences make the trip ends act like fixed constraints. This is the way a trip distribution matrix is conventionally synthesised in a demand model. With weak trip end confidences, the screenline counts act like fixed constraints. This is the basis of many matrix estimation methods.

The ability to meet the EEM and DMRB validation criteria depends heavily on the latitude allowed in trip ends. Without any latitude, the synthetic demand model fails the criteria. With latitude based very roughly on the statistical accuracy of the data the criteria may be met, particularly with the Tanner deterrence function. Matrix estimation may offer a latitude that makes validation easy.

All the Exponential models have the same parameterisation, which is typical for trip distribution but not for matrix estimation. The single cost coefficient is a parsimonious description of interaction effects as a pure trip distribution, while separate factors for every zone is a profligate (even saturated) description of the main generation effects. Typical matrix estimation parameterisation, with one factor corresponding to each screenline count, should be more efficient in meeting validation criteria, but is more limited in the way it can adjust trip ends to suit. (MVESTM provides full trip end parameterisation by default.)

There is a broad similarity in the patterns of GEH values for strong and ordinary trip ends that is not apparent with weak trip ends. This may be because the screenline count estimates are mainly constrained by trip ends when they are strong or ordinary, but when these constraints are relaxed the residual misfitting reflects internal inconsistency between screenline counts.

Table 8.31 shows that total screenline crossings are replicated quite closely. The fit is better with weaker trip ends as would be expected.

Table 8.31 Fit of trip totals – difference from initial values

Trip totals	Initial	Exponential			Tanner
		Strong	Ordinary	Weak	Ordinary
Screenline crossings (3 × hourly)	35,334	2.20%	0.30%	0.01%	-0.08%
Trip ends (24 hour)	191,931	-0.02%	-0.27%	-0.63%	-0.07%

8.7.7.7 Fit to trip ends

Table 8.31 shows that the fit to total trip ends is even closer than that to screenlines, but improves with stronger trip ends. The Tanner function shows a distinctly better fit than the Exponential function for both screenline crossing and trip end totals.

Within this overall fit, the proportional changes in zonal trip ends differ between the sectors defined by the screenlines. Variations within these sectors are much smaller but still show distinct spatial patterns.

8.7.7.7.1 Exclaves

The screenlines do not always follow natural boundaries. In two cases this produces 'exclaves', which are areas separated by major natural barriers from the rest of a sector defined by a screenline.

The count point separating the Hutt Valley from the Wairarapa is not at the natural boundary of the 555m-high Rimutaka pass, but some way down the Hutt Valley, just north of Upper Hutt. This includes five zones at the top of the Hutt Valley in the Wairarapa sector in traffic terms. They are zones 132 to 136 from Timberlea and Brown Owl northwards, including the Akatarawa Valley and Te Marua. This enclave is referred to as Hutt End.

The Pukerua exclave (zones 112 and 113) lies on the Kapiti Coast side of the screenline that separates them from Porirua. Pukerua Bay is separated from the rest of the Kapiti Coast sector by high ground which forces the main transport links into a narrow coastal corridor.

There are unusually large changes to trip ends in these two exclaves. They are shown separately from the rest of their sector in figure 8.8. Although the adjustments to the exclaves are relatively large, the exclaves are quite small in terms of productions and particularly attractions.

Table 8.32 WTSM generations in exclaves – 24-hour HBW by car

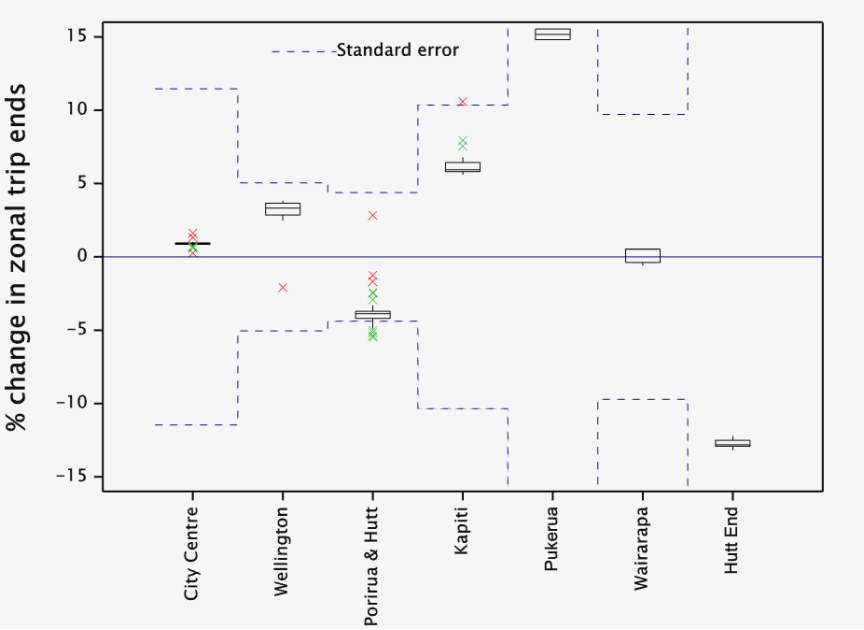
Exclave	Productions	Attractions
Pukerua	862	239
Hutt End	3581	807
Full matrix	191,932	

The consequences for practical modelling are unlikely to be significant. The exclaves include large areas of high undeveloped ground and so appear large in figures 8.9 and 8.10.

These exclaves are differences between the sectors defined by screenlines, and sectors used for aggregation earlier in the study. There are two other areas of difference around Wellington city centre and at Grenada village. Otherwise, the screenlines correspond with the six-sector system of aggregation, with Hutt Valley and Porirua combined into a single sector.

8.7.7.7.2 Differences between sectors

Figure 8.8 Ranges of trip end adjustments, by sector and exclave



Production trip ends fitted with Exponential deterrence function and ordinary trip end confidences

Figure 8.8 shows adjustments to production trip ends when fitting an Exponential trip distribution with all screenline counts separated and ordinary trip end confidences. Box and whisker plots show the range for each sector defined by screenlines. The ranges plotted for the Kapiti Coast and Wairarapa sectors exclude zones in their Pukerua and Hutt End exclaves, which are plotted separately.

The boxes span the interquartile range, so the changes for half the zones lie within the box; the median is shown by a line through the box. The boxes show the narrow central ranges of trip end adjustments within each sector and the wide gaps between sectors. The more extreme adjustments in the Pukerua and Hutt End exclaves are quite distinct from the rest of their parent Kapiti Coast and Wairarapa sectors.

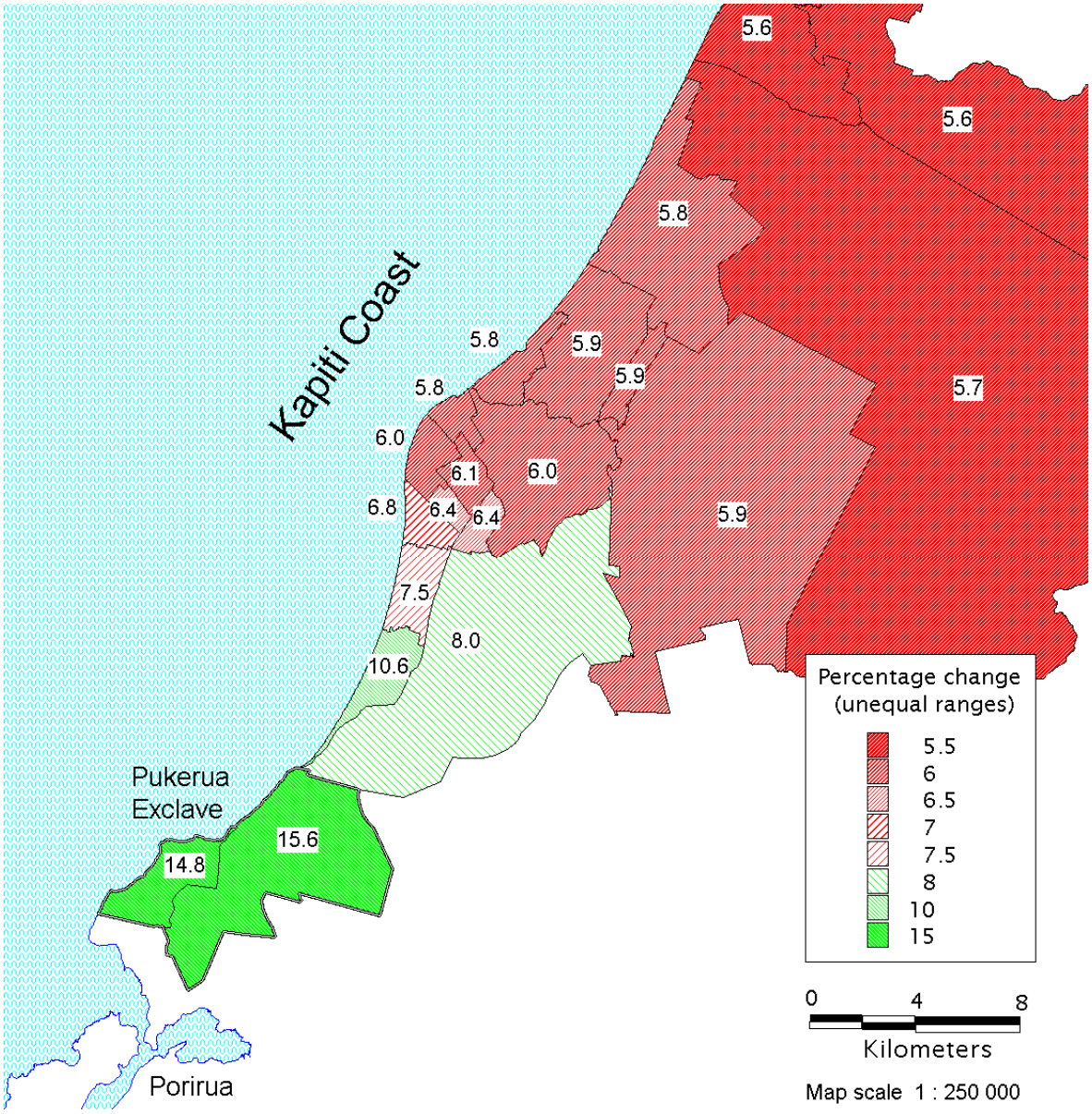
The whiskers extend beyond the boxes to the maximum value, up to a limit of 1.5 times the interquartile range. Beyond that limit outliers are plotted individually. Most of the outliers are zones close to sector boundaries and screenline count points, or in some cases straddling them.

Results for strong and weak confidences and for a Tanner deterrence function show that adjustments are always upwards in the Pukerua exclave and downwards in the Hutt End exclave, for both productions and attractions. The sense of adjustment is generally reversed between productions and attractions for the main sectors.

These major variations in adjustments between sectors are consistent with the theory set out in section 8.5. This shows that the main role of trip end totals in calibrating trip distribution can be seen as infilling segments of the matrix that are not intercepted by screenlines. These infilled segment totals are differences between trip end totals and screenline counts. The differences depend on trip end totals for sectors rather than individual zones within them.

8.7.7.7.3 Differences within sectors

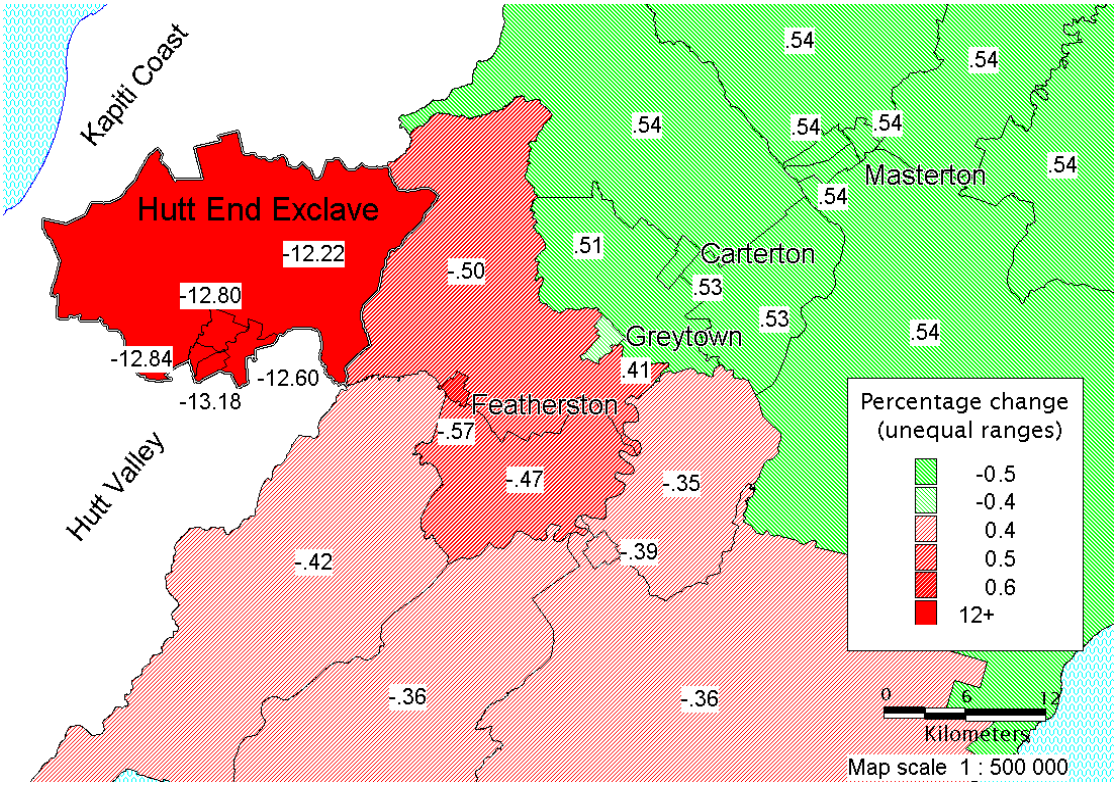
Figure 8.9 Pattern of trip end adjustments – within Kapiti Coast sector



Production trip ends fitted with Exponential deterrence function and ordinary confidences.

Within the narrow ranges in each sector there are distinct spatial patterns. As an example, figure 8.9 shows the pattern for productions along the Kapiti Coast. The adjustments are smallest at the north of the sector and increase very gradually toward the middle of the sector. The increase is then more marked to the south, where there are large adjustments in the Pukerua exclave.

Figure 8.10 Pattern of trip end adjustments – within Wairarapa sector



Production trips ends fitted with Exponential deterrence function and ordinary confidences.

Adjustments in the Wairarapa proper vary by little more than 1%, but still show a distinct spatial pattern within this narrow band. Figure 8.10 shows that adjustments are positive and very consistent to the north of Greytown. To the south the adjustments are negative, decreasing slightly from the south and east towards Featherston. This still differs greatly from the Hutt End exclave on the other side of the Rimutaka pass. The step change at Greytown may reflect some balance of productions and attraction in and around the local centres of Featherston, Greytown, Carterton and Masterton, which lie along the axis of SH2.

Although spatial patterns such as these can be found in most sectors, no general explanation for the patterns is apparent.

The smaller but still systematic variations within sectors are consistent with section 8.5.3, which shows how the within-sector pattern of trip ends is also, but more weakly, related to trip distribution. Apart from a few outliers, trip end adjustments do not appear to be due to routing effects through screenlines.

8.7.7.7.4 Comparison of trip end adjustments with confidences

The adjustments to the trip ends are small and well within the accuracy for individual zones represented by the confidences. However, much the same adjustment applies to all zones within each sector so they are not independent. This needs to be considered in assessing the adjustments applied to whole sectors.

One approach is to sum the trip ends for the whole sector and calculate the standard error for the total trips, assuming a Poisson error process with the trip end confidence:

$$\text{Relative standard error} = 1/\sqrt{(\text{trip total} \times \text{confidence}/100)}$$

These relative standard errors are plotted in figure 8.8. All the mean adjustments fall within them including those for the exclaves, which have wide margins because of their small trip totals.

The deviances in table 8.33 below are scaled by the trip end confidences. The relative reduction in deviance when fitting sector factors is disproportionate to the degrees of freedom, reflecting the relative consistency of adjustments within sectors. However, the absolute reduction in deviance is always less than one per degree of freedom. This suggests that the sector adjustment factors are, if anything, still smaller than would be expected from the scale of the trip end confidences.

The WTSM generation model for HBW attractions (all modes) includes correction factors by eight TLA sectors (TN16.3, table 3.1 and TN16.5, table 3.1). The typical scale of factors is about ±20% and is considerably larger than the adjustments made here with matrix estimation.

The trip end confidences used here are broad estimates with little empirical support, but the trip end adjustments made during matrix estimation do appear to fall well within them. A better understanding of the errors of generation models needs to include their distribution, Poisson-like or otherwise, and spatial correlations, for application at aggregate sector levels.

8.7.7.7.5 Incorporating trip end adjustments in the demand model

The model estimation methods applied here can calibrate a trip distribution model by returning cost coefficients and produce a matrix that fits screenline counts adequately. However, that matrix is not purely a function of the demand model comprising the calibrated trip distribution and the generation model used to calculate the input trip ends. The trip end adjustments have to be included to replicate the estimated matrix that fits the screenlines.

Figures 8.9 and 8.10 show that, strictly, the adjustment factor varies by zone but figure 8.8 shows that much of this variation can be captured by one common factor for each sector. Table 8.33 shows the amount of the residual trip end deviance that can be represented by such factors for different sets of sectors and exclaves.

Table 8.33 Analysis of trip end deviances

Model	Degrees of freedom	Exponential				Tanner			
		Productions		Attractions		Productions		Attractions	
		Deviance	%	Deviance	%	Deviance	%	Deviance	%
Initial	214/223	2.166	~	2.222	~	2.693	~	0.948	~
+ Constant	-1	2.157	0.4	2.213	0.4	2.693	0.0	0.946	0.1
+ Screenline sectors	-4	0.484	78	1.179	47	0.800	70	0.232	76
+ Major exclaves	-2	0.091	96	0.098	96	0.153	94	0.049	95
+ other differences	-2	0.079	96	0.097	96	0.138	95	0.047	95
Aggregation sectors	-5	0.498	77	1.107	50	0.695	74	0.271	71

The deviances in the initial line of the table are residuals from matrix estimation with all screenline, direction and period counts presented separately, and ordinary trip end confidences. They complement the reductions in deviance shown in table 8.26 and the residual fit at screenlines discussed in terms of GEH in the previous section. The initial degrees of freedom for production and attraction trip ends differ according to the number of empty zones omitted, 11 and 2 respectively.

The initial deviances were recalculated from the fit of the trip ends generated by the WTSM to those output from matrix estimation. Ordinary trip end confidences were applied in the deviance calculation. In the following rows of the table, the fit was improved by adding the factors described below. The reductions from the initial deviance are shown as percentages.

- Constant. Fitting a constant adjusts for the differences in total trip ends shown on the lower line of table 8.31. Since these differences are small, the reductions in deviances are also small.

- Screenline sectors. There is a much larger reduction, between a half and three-quarters of the initial deviance when factors are fitted for each of the sectors defined by the screenlines.
- Major exclaves. Most of the remaining deviance is removed by adding separate factors for the two major exclaves of Pukerua and Hutt End.
- Other differences. There is little further reduction with separate factors for the two other areas where the screenline sectors differ from those used in aggregation, around the city centre and at Grenada village.
- Aggregation sectors. This is an alternative set of six sectors, as used to aggregate the observed HIS matrix, following natural boundaries more closely. These give a similar fit to the sectors defined by the screenlines.

In the Exponential model, large adjustments to the attractions of the Pukerua exclave limit the fit of either set of sectors. The adjustments are around +90%, but the standard error is about the same because of the small total of attraction trip ends, 239. The Tanner function makes much smaller adjustments to the attractions, at the expense of larger adjustments in the productions.

A formulation of MVESTM to constrain trip end adjustments to common sector factors has been envisaged. In the limiting case of only one sector, if zonal planning data is entered in place of trip ends, the resulting single fitted factor is the trip generation rate, eg trips per household. This is joint estimation of the generation and distribution models, as considered in section 8.5.1.2.

The formulation is complex and has not been tested. Given the large proportion of zonal trip end adjustment that can be described by sector factors, it seems likely that there would still be an adequate fit to screenline counts. This is even more likely if exclaves are allowed separate factors, or better still avoided in the study design.

Multiple sector factors act like the K factors in trip distribution, but for the generation models. They are subject to the same arguments about the inclusion of arbitrary empirical factors. Adjusting trip ends by sector might be seen as transferring a trip distribution problem to the generation stage. However, such adjustment:

- is within the accuracy of the generations, as well as it can be assessed
- is a concise and parsimonious model that is already applied in the WTSM to attractions
- follows good statistical practice in modelling main effects before interactions.

The insensitivity of cost coefficients to trip end confidences across the bottom of table 8.29 argues against a fundamental problem for the trip distribution model from trip end adjustments.

The method will work best with sectors that are clearly defined by screenlines. Sectors will be harder to define around incomplete screenlines, particularly if screenlines are subdivided and presented as their component link counts. However, this study has found distinct sector factors given the practicalities of screenline location and sector definition in the working WTSM model.

8.8 Discussion

Trip distribution model calibration from aggregate (count) data has been explored empirically in a Wellington model (WTSM) using the MVESTM software package. Trip end information has emerged as important and theoretical bases for needing trip end data have been offered in section 8.5. The empirical development has addressed some practical issues, in particular factoring between all-day production-attraction matrices and counts observed by direction and period, but has artificially avoided others, being

restricted to commuting trips by car. This section discusses issues arising from the research but not addressed directly by it, particularly issues affecting its wider practical application.

8.8.1 Issues not addressed

The empirical analysis has worked within or sometimes around the constraints of the MVESTM programme. These reflect the availability of computer core memory when the code was written in the 1990s and its usual application for matrix updating. As has been demonstrated, it is a much more flexible tool allowing separate specification of model structure and data. Some extensions have been formulated within MVESTM's current limits but there are still greater possibilities for the methodology underpinning MVESTM, ie fitting a log-linear model by maximum likelihood of Poisson-distributed data. This is the same as the GLMs fitted to disaggregate data using GENSTAT in previous chapters and the methods developed there point to the possibilities for model fitting from aggregate count data.

The principal advantage of calibration from aggregate data is that counts are abundant and simple to collect. Trip distributions are usually calibrated from matrices observed by intrusive and expensive interview surveys. However, the data from such surveys, in particular HIS, is also used to calibrate other stages in the model and to derive other useful model parameters. If the costs and difficulties of interview surveys are to be avoided, another source of information for these other model components is needed. This might again be from aggregate count data, an alternative survey, or some combination as in the use of trip end information when calibrating the trip distribution model from counts. Possibilities for using aggregate data to calibrate log-linear models are discussed below.

Counts from four screenlines in Wellington have shown considerable power to detect trip distribution effects and determine the deterrence function, equivalent to about half that of the 2500 household interview survey. Even if the calibration/fitting of other model components can be formulated in MVESTM or as a more general log-linear model, it is hard to say whether count data can provide sufficient information in practical cases without empirical evidence. Theoretical approaches such as those set out for simple cases in section 8.5 might be extended to practical cases but not without difficulty.

8.8.1.1 Dependency on trip ends

In Wellington, contrasts between the screenline counts alone provide little information about distribution and calibration depends on contrasts with trip end totals. Section 8.5 argues that this is the general case in practice but also shows that planning data (eg numbers of homes and workplaces) can be used as a proxy for trip ends.

Joint fitting of generation and distribution, ie simultaneous calibration of trip rates and deterrence function, has been formulated for MVESTM for both uni- and multi-variate generation models. The multivariate formulation may address the issues identified by Daly (1982) for attraction variables.

Many of the cases of model estimation reviewed in section 8.1.2 took this approach of calibrating a joint generation-distribution model, using planning data as well as link counts. All of the practical exercises used some such information about trip ends; none calibrated a real trip distribution from link counts alone.

8.8.1.2 Fitting period and direction factors

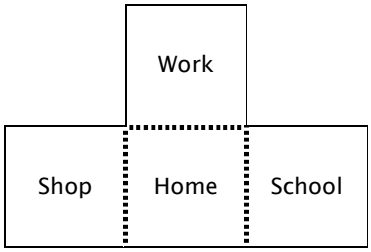
This study has addressed the use of counts by period and direction by incorporating fixed factors derived from external sources (section 8.6.3.3). It appears possible to fit these direction and period factors from count data using a similar approach to that formulated for joint calibration of generation and distribution.

8.8.1.3 Multiple purposes

This study has calibrated trip distribution models for a single purpose (HBW) only. Transport models are usually stratified by purpose (Brown 1982), which cannot be distinguished in vehicle or passenger counts. Counts will

be aggregates over all purposes as well as over the sets of OD movements intercepted at the count point. Different purposes are usually modelled by different trip distributions with different deterrence functions and different trip ends representing the different locations of homes, workplaces, schools and shops.

It appears theoretically possible for several purpose-specific distributions to be estimated simultaneously, if, for instance, a model area comprised distinct sectors for housing, employment, education and shopping with screenlines separating the sectors. Counts on those screenlines would control the total of each purpose-specific trip matrix; further screenlines would be needed for trip length information to calibrate their distributions.



In practice, land uses within sectors will be much more mixed. Sectors are defined by screenlines. In a large regional model, good screenlines, with few crossings and unambiguous movements, lie on natural boundaries. The communities on either side tend to be self-sufficient with their own balanced mix of land uses. The Rimutaka and regional screenlines are examples in this study in the Wellington region.

A central area cordon might separate residential productions outside from attractions inside. The movement between them would give a control total of trips of all home-based purposes, but no distinction between them.

However, cordon counts can also intercept through traffic and in a complex model with multiple screenlines, a screenline might intercept movements from several sectors on either side. Any distinctions between purposes arising from concentrations of land uses in particular sectors will be diluted in the aggregation of counts across the different sectors.

Four major screenlines have been analysed here. Assuming the full set of counts by period and direction effectively yield two distinct sets of PA movements, one in each direction across the screenline, this gives eight items of information. This is just sufficient to calibrate either four purpose-specific trip distributions which need both scale and a deterrence parameter, or eight purposes if the scale is determined by trip ends. However, given the considerations above, it seems unlikely that screenline counts will carry much distinctive information for either case. Further clear screenlines become progressively harder to find and possibly less informative if land uses on either side are similar. The opposite case would be tight cordons around specific land uses, such as industrial areas or large shopping centres, though their information might be used more simply to calculate trip rates for the generation model.

Individual purposes also need their own period and direction factors. Commuting is usually strongly tidal, with most journeys to work made during the morning peak and returns during the evening peak. Education trips may share this pattern, but the patterns of other purposes such as shopping and business may be much less distinctive. It seems most unlikely that a multiple-purpose demand model can be calibrated completely from aggregate count data and zonal land use data alone. Even with trip rate, period and direction factors provided, practical screenline counts seem to offer little information to distinguish deterrence parameters by purpose.

Low (1972) offers a simple approach to scaling the generations of different purposes with fixed trip distributions. Cascetta and Russo (1997) re-calibrated their model for two purposes simultaneously, but did not show the power of information available from a practical network.

MVESTM cannot be readily formulated to fit multiple purposes with separate deterrence functions because it incorporates only one set of cost parameters, alpha and beta. This is a limitation of the MVESTM software package but not of the log-linear models it fits.

8.8.1.4 Multiple modes

These estimation methods can be applied to public transport models. Ticketing information may provide station-to-station volumes as well as simple link counts which can be stratified by service. However, all these can be interpreted as aggregations of proportions (linear combinations) of OD matrix cells which can be determined from an assignment model.

Public transport and highway counts are usually distinct, so trip distributions can be calibrated for each mode separately.

A single all-mode trip distribution could also be calibrated from the count data. Adjustments are needed between passenger counts on public transport and vehicle counts on highways. This study has already incorporated car occupancy in the intercept proportions (section 8.6.3.3).

Where screenlines are strongly defined by a common barrier such as a river, public transport and highway counts may intercept the same set of OD movements and the counts can simply be added together. However, the scopes of public transport and highway counts will usually differ, particularly with limited-stop transit services and limited access highways.

A mode split model can estimate a modal share for each cell in a trip matrix. These shares could be incorporated in the intercept proportions that define the scope of a count, in the same way that period and direction factors have been incorporated in this study.

In theory, a common trip distribution model could be calibrated from counts of just one mode. However, this would place a heavy reliance on the modal split model, probably more so than the reliance on the assignment model in conventional matrix estimation. Tamin et al (2003) calibrated a mode split and trip distribution model from counts of only public transport passengers.

8.8.1.5 Multiple model stages and levels in choice hierarchy

It is possible to formulate a joint calibration of mode split and distribution, or of other combinations of model stages.

However, the log-linear representation of choice outcomes does not allow a complete separation between the cost coefficients for the different dimensions of choice, eg mode and destination. The relative size of these coefficients determines the proper order for the stages in model building to avoid perverse elasticities (see section 2.5.1).

The differences in cost coefficients can be seen as different levels of randomness in the utility of choices. Greater certainty in some choices can be interpreted as a correlation between them, which can be represented in a mixed logit model.

Chapter 6 sets out the hypothesis that mixed logit models can be fitted by hierarchical GLMs, an extension of log-linear models. However, it fails to confirm this hypothesis empirically, so it may be very difficult to develop an algorithm for joint fitting along these lines even if the hypothesis is true.

Section 6.4.4.3 suggests that a large amount of information is needed to distinguish the correlation effects that result in separate cost coefficients. On the other hand, the analyses in this chapter show that

screenline counts have considerable power to calibrate trip distribution and since counts are distinguished by mode they may well have considerable power to calibrate mode choice too.

In their joint estimations, Cascetta and Russo (1997) re-fitted separate coefficients of cost for mode split and for trip distribution, but Tamin et al (2003) did not report a cost coefficient for mode split.

Modelling trip generation from land uses has already been discussed as a vital complement to trip distribution modelling in the absence of zonal trip ends. Hierarchy of choice is not an issue unless there is cost elasticity, ie an accessibility term.

Time-of-day choice is similar to mode choice in that time of day can be readily distinguished, like mode, in most surveys, whereas origin, destination and purpose generally remain latent in counts. Unlike different modes, different time periods will usually be observed by the same survey. There are likely to be correlations between adjacent time periods just as there tend to be correlations between certain modes, eg public transport.

An inability of log-linear models to represent such choice hierarchy may be a fundamental limit to calibrating the best current forms of transport model from aggregate data.

8.8.1.6 Multiple cost components

In principle, aggregate count data can be fitted to a log-linear model that distinguishes multiple components of cost such as time, distance and tolls or fares. MVESTM's cost parameters alpha and beta can be seen as coefficients of two components, log(cost) and cost respectively.

In practice, separate time and distance components have been fitted by GLM (section 4.10), but showed relatively little significance even in the HIS data at the zonal level. To distinguish such components using counts would not only require some destinations to be closer in time while others are relatively closer in distance, but also for such distinctions to be maintained in the aggregation of OD flows to link counts.

Costs in different model stages may have different components and these may be even more difficult to distinguish in aggregate data. For example, a link count will be affected by the generalised cost of assignment as well as the generalised cost of trip distribution.

8.8.1.7 Alternative error distributions

The log-linear model in MVESTM is fitted by maximum likelihood of Poisson-like errors in the data. A Poisson distribution is expected from the sampling of the HIS, but there is no such prime facie expectation for the characteristic distribution of link counts or synthesised trip ends. Section 8.4.4 shows some support for a Poisson distribution of link counts and a gamma distribution of trip ends, but there is little clear evidence as to the scale of such errors. A Poisson-like error with variance proportional to the mean was adopted thoughout the empirical analysis for simplicity.

If the error in some data has a known distribution other than the Poisson, it can still be represented in MVESTM to a first order approximation. To do so, MVESTM's confidences are set such that the expected variance of each item of data is correctly replicated within MVESTM, where:

$$\text{variance} = (100/\text{confidence}) \times \text{mean}$$

For example, if errors follow a gamma distribution, where the standard deviation is proportional to the mean, with a coefficient of variation k

$$\text{standard deviation} = k \times \text{mean}$$

then confidences are then calculated as

$$\text{confidence} = 100/(k^2 \times \text{mean})$$

The confidence entered into MVESTM thus depends on each observed value of the mean, even for data of the same type and provenance. This approach was applied by Gunn et al (1997, appendix A.3) for dealing with non-Poisson group observations in The Netherlands.

It may be formally correct to iteratively re-calculate the confidence to maintain the expected variance as a function of the fitted mean, like the iterative re-weighting used to fit GLMs. However, such sophistication cannot be justified by the understanding of errors achieved in this study. The re-calculation would be needed where observed and fitted means differ substantially; such differences are themselves more worthy of attention. Correct weighting minimises estimation errors in a well-specified model, but cannot of itself address errors of specification.

GLMs can accommodate many different distributions, but the Poisson may pose fewer problems in fitting the log-linear form because it is conjugate with the logarithmic link.

8.8.1.8 More information from individual link counts

The number of items of data can be increased by presenting screenline counts as their component link counts at the individual crossing points. However, this introduces assignment errors and is a rationale for the wider validation criteria in the EEM and DMRB for links than for screenlines.

There may be some increase in the power of discrimination in the data if there is some variety in the movements served by the different links. For example, local roads may carry shorter trips or serve a particular land use nearby.

The accuracy of synthesised trip ends is likely to limit the improvement in accuracy from more count data as shown for a simple case in figure 8.3. Another indicator is the information, as measured by reduction in deviance with the introduction of a cost term, available from different aggregations of the observed matrix in table 8.15. This shows that trip totals for 36 segments between six sectors offer three-quarters of the information available from the full zonal matrix. Section 8.7.5 shows that the four segments roughly defined by the radial screenline offer a quarter of this zonal information.

It may be useful to consider the limiting case of a count for every link in a network. This cannot guarantee an exact fit for a trip distribution model and is likely to violate continuity conditions at nodes, which will impose a degree of redundancy on the dataset. The errors and uncertainties which remain in this dataset may show diminishing returns for more link counts.

8.8.1.9 Multiple data sources

The MVESTM programme is effectively limited to accept just one matrix of trips or costs. This study has circumvented the restriction by entering components of a trip matrix as individual screenline counts. More than one matrix of trip data could be entered thus.

For example, there might be one observed matrix from each census station in an RSI. This offers to incorporate matrix compilation in the model calibration. Such matrix compilation is usually a separate process of data preparation but can imply or be formally structured around a statistical model as in the UK Department for Transport's ERICA software suite developed by Peter Davidson.

Basic matrix building is contrasted with model and matrix estimation in section 8.8.5.

8.8.1.9.1 Multiple levels of aggregation and disaggregation

Observed matrices are actually disaggregate to the zonal level. Potentially any data that is a known linear combination of matrix cells can be considered in the fitting of a log-linear model. Usually in matrix estimation the data is link counts, aggregated over the OD movements which use the link, but turn counts can also be used and the practice in MVESTM is to aggregate to screenline counts. In this study, the observed matrix has been aggregated to segments of different sizes. Trip end totals can be seen as aggregations of

matrix rows and columns. Part-route data from numberplate matching and station-to-station ticketing can be described as an aggregation of matrix cells and incorporated in the model calibration.

The model may be generalised to further dimensions such as purpose, mode, time period or car availability. These dimensions can be seen as factors in a GLM, giving stacked or hyper-matrices in transport modelling terms or tensors in mathematical terms. Data may then be aggregated across dimensions such as purpose, or confined within one level of a dimension such as mode.

The model might be disaggregated from production zones to households, persons and trips. This was done with GLMs in chapter 5, allowing the effects of variables specific to household, person or trip to be examined. All the other data considered above could still be described as aggregates across this model structure. Cascetta and Russo (1997) offer formulas for combining disaggregate data with link counts. It may need a lot of computing power.

The method has the potential for combining information from many sets of data at different levels of aggregation.

8.8.1.9.2 Model adjustment

This study and papers by Tamin have found the need for good starting values for model parameters. Link count data may not have a lot of information to determine some parameters, such as deterrence functions by purpose.

Rather than try to calibrate a full model from count data alone, it may be better to use these methods to adjust an existing model or to test the need for adjustments.

A demand model could be synthesised with separate purposes, with initial parameters from an earlier study, another study area or a small disaggregate survey. It would be introduced as a fixed prior matrix in matrix estimation terms or an offset in a GLM. This would form the base of the log-linear model structure, but would not (necessarily) be multiplied by a fitted coefficient or treated as data for the final model to match as is the default for a prior trip matrix in MVESTM.

One coefficient would be fitted to all costs. The coefficient would then be applied as a common adjustment in the deterrence function for each purpose. If the purpose matrices were presented separately within the demand model (eg as submatrices of a hypermatrix), the method could test for the significance of any distinctions between purposes in the aggregate data.

8.8.2 Application

8.8.2.1 Calibration

It seems doubtful that relatively cheap count data would provide sufficient information to calibrate the whole of a full transportation model without resort other sources. The weight of information from counts might reduce the need for expensive interview surveys on a large scale and the general methods developed here might allow both sets of data to be included in a single calibration. This is very similar to using revealed preference data for the overall adjustment of a stated preference model and the method might be used for that purpose.

In smaller studies, local count data might be used to check or adjust parameters in an existing demand model. The existing model might have been calibrated previously; this is then similar to conventional matrix estimation, except that the objective is to update demand model parameters rather than the matrix. The initial demand model may be built with default settings in ‘off-the-shelf’ packages or generic values prescribed in advice manuals. This is similar to comparing local accident records with national accident rates; model fitting may be able to abstract more from local data than a simple comparison of

flows. Ideally, the accuracy of the old or generic model would be known so the final model could represent the best fit between it and the new or local count data.

8.8.2.2 Validation

The main value in calibrating a demand model with link count data (if not exclusively from it) may be in meeting the validation criteria for those counts. This will allow the matrix from the demand model to be used directly in the assignment model, obviating the need for awkward and sometimes coarse linkages between demand and assignment models.

Apart from the computational convenience and clarity in running the full model, a formal analytical bridge between demand and supply modelling allows the demand model to be examined for causes behind poor validation on the network and allowance made for uncertainties in it.

The demand model may need some empirical adjustment for validation. In this study, zonal adjustment factors were fitted for trip ends. The adjustments fell within the (very broadly estimated) accuracy of the generation model and it appeared that validation could be achieved with a more parsimonious set of sector adjustment factors, such as those already incorporated in the WTSM generation model. Validation of a full demand model that is more extensive than this study's single purpose, single-mode trip distribution might be achieved by spatial or temporal variation in other factors. The analytical linkage between the demand model and validation counts allows formal specification and testing of alternative adjustments.

As with any fitted coefficients, but particularly empirical ones, there are questions of model specification and temporal stability. These are no worse for demand model adjustments than for equivalent adjustment factors applied in matrix estimation of a base assignment matrix, or implicit in its linkage with the demand model. Demand model adjustments are, by design, likely to be simpler, more parsimonious and more readily comprehensible.

8.8.2.3 Calibration with validation data

There is also a question as to the propriety of including validation counts in a calibration. Retaining some data from model fitting as an independent check is a respected practice; there are sophisticated methods which can be very robust. However, maximum likelihood models offer a strong statistical framework for measuring and interpreting their fit to data without excluding any of it from the calibration process. Including validation data in the calibration can artificially exaggerate the model's fit to that data; this lack of independence can be compensated in tests by adjusting degrees of freedom. Statistical measures of fit are not all as readily comprehended as link count comparisons, although the GEH can be similar to the deviance. Statistics of fit are specialised tools, like economists' elasticities.

Link counts may be retained for independent validation simply because they are not readily incorporated in the calibration of a demand model without the methods employed in this study.

There is a strong argument for using all data available to provide the most accurate model possible. If count data is retained exclusively for validation, the cost of collecting sufficient interview data for calibration may be much higher.

A good fit between modelled and observed link counts is not just a convenient measure of a model's quality. It is an important property of a model to replicate flows in a corridor or across a barrier where investment in new facilities is being evaluated.

Validation may be falsely improved by overweighting the validation data during calibration. The internal fit of the calibration model offers a check on this: after allowance for weighting, the fit of validation data should be no better than the fit of the other data.

8.8.3 Alternative computational approaches

All the full-scale analysis was performed with MVESTM, taking advantage of its ability to optimise parameters of a deterrence function based on a cost matrix. The following possible computational approaches that do not depend on this specific facility emerged during the study.

8.8.3.1 Trial and error

Since some deterrence functions are defined by a single parameter, it can be quite practical to search for its best fitting value by successive approximation. A deterrence matrix could be formed from costs with a trial value of the parameter and Furnessed to fixed trip ends. The resulting synthetic trip distribution would be compared with observations by an objective function. The objective function could be a maximum likelihood as in GLMs and MVESTM, or another measure of fit.

The search for an optimum becomes more difficult very quickly as the number of parameters increases. The Tanner deterrence function requires a second parameter; K or L factors or sector trip end adjustments will usually require more. Allowing adjustment of zonal trip ends in accordance with their confidences requires a number of parameters that are twice the number of zones, less one.

Modern computing power can speed up the process, but will usually be more effective in algorithms that recognise the structure of the problem and take advantage of its properties in searching for a solution.

'What-if' tables in Microsoft Excel were used in an initial manual search for a minimum deviance, and the Solver add-in found the root of the cubics for figure 8.2. Genstat was used as a brute force optimiser for this and to optimise two parameters for the relative confidences in figure 8.3.

8.8.3.2 Cost bands as screenlines

A matrix estimation algorithm may be adapted to calibrate an empirical trip deterrence function by specifying a parameter X_k for each cost band. The corresponding intercept proportion R_{ijk} is set to 1 if the cost of movement from i to j falls into band k, and zero otherwise. The empirical deterrence function is calibrated by the set of fitted parameters.

Conceptually, if drivers pay the cost of their trip at toll booths, with a specific toll booth for each cost band, the traffic through each toll booth is intercepted by a screenline.

Unless there is information about the total number of trips falling within a cost band, the algorithm needs to accept a null return. Conversely, the parameters corresponding with actual count data need to be suppressed, or they will risk aliasing with the deterrence function. Finally, trip end data and corresponding parameters are needed, completing the tri-proportional structure of the empirical trip distribution model.

Empirical deterrence models can be profligate in degrees of freedom, taking up one for each cost band after the first. Aggregate count data may provide only one or two degrees of freedom per screenline; the four Wellington screenlines do not support a second parameter in the Tanner deterrence function.

Cost bands were used by Gunn et al (1997) in their estimation of a Dutch national matrix with HCGMAT, and by Bell et al (c2006) in their path flow estimation on the Swiss national network. However, their purpose appears to have been to introduce external trip length information as a constraint, rather than to calibrate a deterrence function. The existence of a control total (or proportion) for each cost band would overcome the risk of indeterminacy. The surveys from which such cost band totals are derived may offer more information about trip distribution at a more disaggregate level.

8.8.3.3 Costs as intercepts

Analytical deterrence functions are more parsimonious in degrees of freedom than empirical functions. Analytical forms may be calibrated from a single parameter X_k of a matrix estimation by transforming the intercept proportion R_{ijk} to represent costs.

Omitting subscripts, the factor X^R in the matrix estimation model can be re-written

$X^R = \exp(\ln(X^R))$

$= \exp(R.\ln(X))$ cf $\exp(-\lambda.Cost)$ where Cost = R and $\lambda = -\ln(X)$

$= (\exp(R))^{\ln(X)}$ cf $Cost^{-\gamma}$ where Cost = exp(R) and $\gamma = -\ln(X)$

Thus:

- for an Exponential deterrence function, set intercept $R_{ijk} = Cost_{ij}$
- for a Power deterrence function, set intercept $R_{ijk} = Ln(Cost_{ij})$

In either case, the cost coefficient is minus the log of the fitted parameter X_k

This transform is a continuous form of the binary 0,1 coding of R_{ijk} to represent cost bands in an empirical deterrence function. For the Exponential, the equivalent 'count' data is the total travel cost, which is the sum over all matrix cells of trips multiplied by the intercept, cost.

Matrix estimation allows multiple screenlines, indexed by k. Multiple 'screenlines' can introduce different formulations of cost to calibrate more sophisticated distribution models, such as:

- the Tanner deterrence function, from cost and Ln(cost)
- separate components of cost, eg time and distance
- L factors for different segments of the matrix, or
- separate coefficients for different purposes, in slices of a hypermatrix.

A brief trial with MVESTM and the WTSM HIS observed matrix gave satisfactory results for an Exponential deterrence but encountered negativity problems with the Power function.

It is quite likely that this transform or an equivalent process is performed in the core of MVESTM to fit its cost parameters alpha and beta.

8.8.4 Software

With the above transformations, potentially any matrix estimation package can be used to calibrate trip distribution models from counts. However it still needs:

- flexibility to separate data and structure
- ability to manipulate intercept proportions to describe both
- a robust algorithm, to accept real numbers as intercept proportions, including negative values and not just in the range 0 to1, and to fit parameters with wide ranges.

For practical use, a good system should offer:

- a good structure, recognising and working easily in dimensions such as car availability, purpose, mode and time as well as origin and destination (eg OmniTrans' Cube)
- a similar ability to handle different levels of aggregation (eg EMME/2 groups) and disaggregation – household, person, or trip

- a structure relating OD and PA matrices, and daily and hourly flows, with period and direction factors
- acceptance of multiple data sources, eg individual RSI site matrices, or components of cost (time and distance)
- a clear distinction between observed zero cells and null cells not observable.

As an analytical tool, a package that fits a log-linear model by maximum likelihood could offer:

- omission of empty zones and other null data from model fitting
- acceptance of an offset matrix or parameters to form a fixed part of the model structure without necessarily being treated as data to which the fit should be optimised.
- flexibility in error distributions; normal and gamma as well as Poisson.
- a full description of model fit, ie:
  - likelihood of overall fit, with attribution to individual components
  - estimated parameters, their standard errors and correlations
- predictions and their accuracies for:
  - individual cells of the output matrix
  - data inputs
  - model structures, eg trip ends or K factors; in simple matrix estimation these correspond with the data inputs
  - other linear combinations of output matrix cells, eg representing patronage or benefits of a new scheme
- measures of validation against observed data, eg GEH for screenlines:
- summary aggregations of these statistics, eg
  - by type:
    - prior matrix or observed matrices
    - origin trip ends
    - destination trip ends
    - screenline counts
  - by location – central, inner and outer cordons
  - by source – boarding/alighting counts, on board counts, ticketing
  - by dimensions such as mode or period
  - by user defined groupings
  - and in particular, contrasting validation data with other data.

These include statistical tools for checking and assessing GLM models which can be applied to this model estimation. Proper interpretation of these measures needs the real accuracy of input data. Relative accuracies may be re-scaled by the overall fit of the model.

8.8.5 Model estimation, matrix estimation and matrix building

This study into model estimation has used techniques designed for matrix estimation and in discussion has touched on matrix building from interview surveys. All three can be approached as a statistical estimation process of fitting a log-linear model by maximum likelihood, which offers a common framework.

A modelling approach to matrix building has been taken in the UK Department for Transport's ERICA software suite developed by Peter Davidson, in the European MYSTIC project (IVT Heilbronn & Sandman Consultants Ltd 1999; Gaudry 1999), and for the Dutch base matrices by Gunn et al (1997). However, matrix building and matrix estimation are usually distinct processes in practice. Table 8.34 outlines differences between their typical basic usages and the trip distribution model estimation developed here.

Table 8.34 Features of model estimation, matrix estimation and model building

Feature	Model estimation	Matrix estimation	Matrix building
Output	Model parameter	OD matrix	Matrix
Prime data	Counts	Counts	Interviews
Other data	Cost matrix, trip ends	Prior trip matrix	
Output matrix	Smooth	As prior matrix	Lumpy
Intercepts used	In program	In program	Manually
Structure	Theoretical	Empirical	Saturated
Parameterisation	Parsimonious	Profligate	Maximal
Model form	$T_{ij} = a(i)\, b(j)\, f(c_{ij})$	$T_{ij} = t_{ij}\, \Pi_k X_k^{R_{ijk}}$, t_{ij} is prior	T_{ij} = mean(t_{ij}), t_{ij} are observed
Notes	a(i), b(j) and $f(c_{ij})$ are special cases of $X_k^{R_{ijk}}$ a(i) and b(j) mutually exclusive, and orthogonal. K and L factors may be added by design	scopes of R_{ijk} are arbitrarily determined by count locations, overlapping	

8.8.5.1 Output

The objective of model estimation is the calibration of a model, specifically finding the value of certain parameters; a fitted matrix is a by-product. In matrix building and estimation, the final matrix is the main product, and any other parameters fitted in the process are incidental.

The interview data used in model building usually includes the purpose at both ends of the trip, allowing either an OD or a PA matrix to be built. The direction of counts is related only to origin and destination, so estimated matrices have to be by origin and destination, more closely related to network flows than to travel demand arising from land uses.

8.8.5.2 Model structure and form

The underlying model structures differ to suit the three processes, and affect their properties and problems. Model estimation is based on the trip distribution model, with theoretical bases in maximum entropy and random utility, providing a parsimonious description based on as little as one key parameter, the cost coefficient. Matrix estimation adopts an empirical model based on the data with one parameter corresponding to each count. Matrix building has a saturated model of the matrix, in effect finding a separate parameter for each cell.

This saturation precludes any extrapolation from observable cells to other, null cells, leaving the need to infill observed matrices built from roadside interviews. This can be achieved by partial matrix methods that exploit the gravity model's generality.

The gravity model's ability to estimate trips in all cells from all the data also produces a smooth synthetic matrix, without zeros except for empty zones. Since there are usually few interviews per cell, individual cells in an observed matrix are subject to sampling error with a Poisson-like distribution producing a lumpy matrix with many zeros. Matrix estimation propagates the characteristics of the prior matrix.

The distinction between unobservable cells and observable cells with a sample of zero is a vital one in building matrices from roadside interviews, or calibrating a gravity model from them. An obverse of the problem of unobservable cells, whose movements are not intercepted at any interview site, is double- or multiple-counting of cells whose movements pass through more than one interview site.

8.8.5.3 Intercepts

It is a good practice to tackle this at the survey design stage. Complete, robust screenlines of interview sites are formed to divide the study area into distinct sectors. For each corresponding matrix segment, a number of screenlines that intercept the movements can be defined. That number will be zero for unobservable movements, within sectors between screenlines. Where the number is more than one, it divides the count of interviews from all the screenlines to correct for multiple counting.

Where a screenline only intercepts part of a movement and another screenline intercepts the whole movement, interviews from the first screenline may be omitted for simplicity. Alternatively, abutting partial screenlines may be combined to provide complete (once and only once) interception. A clear view of which movements should be intercepted at each screenline is of great value in checking interview records.

This approach of clarity in design allows manual correction for multiple counting, which may be applied by manipulation of interview records alongside sample expansion, or of built matrices. It avoids routing issues. The main statistical issue is weighting between surveys with different sampling rates.

Matrix estimation and hence the model estimation studied here also has to relate its primary data, aggregate counts, to the sets of movements they comprise. However, the scope of movements intercepted in a single count is not generally a 'rectangular' matrix segment neatly defined by sectors and there will be fractional intercepts under multi-routing. Methods for matrix estimation have developed to handle such complex intercept patterns within their computer algorithms, as well as fitting to data aggregated over them. The algorithms similarly account for arbitrary overlapping and repeated counting but do not recognise complementary counts (eg along a screenline) that form one logically complete intercept, free from routing error.

Gunn et al (1997, appendix A.3.2) show how overlap between observations determines the gradient of the Poisson likelihood with respect to the parameters fitted in conventional matrix estimation. Again, if two count points, k and l, both intercept half the traffic from i to j ($R_{ijk} = ½$, $R_{ijl} = ½$) the derivation does not distinguish between the same half being counted twice in series along the same route ($R_{ijk} \equiv R_{ijl}$), or complementary halves being counted on parallel routes ($R_{ijk} \not\equiv R_{ijl}$). The formulation suggests independence, assuming a quarter of the i to j trips are counted at both sites k and l. A path-based derivation may be illuminating.

Complex intercept patterns can be handled in matrix building by using the select link matrix for the interview site. This can be used for interview record checking or for sample expansion by the reciprocal of the intercept fraction. Alternatively, interviews can be treated as single cell counts in a matrix or model estimation.

8.8.5.4 Model parameterisation – K factors

The saturated model structure for matrix building, which gives complete independence between cells, allows any pattern of matrix to be represented. The parsimonious gravity model allows only limited sets of matrix patterns by distributing trip ends in according to the deterrence function of costs.

This sometimes appears too restrictive to represent actual travel patterns and extra parameters are introduced to improve the model's fit. These are K factors which are constants in the deterrence function that vary between matrix segment or L factors which vary the cost coefficient. K and L factors can be segmented as in:

- London – simple spatial – cross-river, central/inner/outer/external sectors
- Wellington – spatial hierarchy
- Scotland – internal Edinburgh, internal Glasgow and other, or
- Dublin – socio-economic linkages.

Usually the segmentation for these factors is by design, remaining quite parsimonious with mutually exclusive segments. They are factors in the GLM sense, ie equivalent to one dummy variable per segment. They may be based on cordons and screenlines where there is data pointing to the need to improve the fit of a simple gravity model.

The parameterisation of matrix estimation can be seen as a set of K factors designed to give an efficient fit to the counts. There is one parameter per count, usually many more than in a set of K factors. Instead of an integer 0,1 dummy variable for each exclusive segment, the variable is the scope of movements intercepted, irregular and with fractional values for multi-routing. These scopes will usually overlap in complex patterns except where there is careful design in the count information, eg counting on complete, robust screenlines and presenting the counts as screenline totals. In this case, the parameterisation can become similar to a conventional set of K factors.

The maximal model for matrix building can be seen as fitting a separate K factor for every cell.

8.8.5.5 Absorption and representation of cost effects

The implicit design of matrix estimation parameterisation to fit the counts will absorb any effects of cost deterrence in the count data. Trip distribution effects are also likely be represented in the prior matrix, particularly if it has been synthesised. Fitting cost as an additional variable in a matrix estimation is unlikely to calibrate a valid cost coefficient for trip distribution or even a significant one.

Most sets of K factors will absorb some effects of cost deterrence. Geographically based segmentations (and particularly the WTSM's hierarchy) tend to stratify by trip length, if only because intrasector trips are shorter and hence will have some correlation with cost. Because cost is a continuous variate, any orthogonal variable must also be continuous so a simple set of K factors cannot be orthogonal and must be aliased with cost to some extent.

When K factors are incorporated in a gravity model used for forecasting, it implies the factors represent a real effect in the base year that will remain the same in the future year. The factors may include some effects of cost deterrence but these will not respond to changes in cost in the future year.

In the limit, matrix estimation can be similarly used for forecasting. The parameter corresponding to each count has to be saved and applied to the same scope in the future like a K factor. There can be little or no residual sensitivity to changes in travel cost; the base year pattern of trip distribution will be frozen in a form determined by the scopes of counts and the prior matrix. Forecasting becomes limited to the effect

of changes in trip ends, which can be achieved by Furnessing the estimated matrix. This preserves its deterrence pattern.

Forecasting with matrix estimation parameters is essentially the same as the matrix adjustment used in the WTSM. However, the base year matrix estimation took a synthetic prior matrix from the demand model. The demand model was rebuilt on future year costs before matrix adjustment with the estimated parameters, so the cost sensitivities of the demand model are represented in forecasting.

K factors can be fitted in model estimation from counts; they can be formulated within MVESTM if their segmentation can be described in its intercept file. However, there is a particular risk of aliasing with cost information if the segmentation corresponds with screenlines of the counts being used to calibrate the cost coefficient. For example, the cost effects derived in section 8.7.3 from aggregate counts for all segments would be completely aliased with a set of K factors for those segments.

Log-linear models fitted by maximum likelihood provide a common framework for these diverse aspects of matrix formation, allowing combinations and permutations of the different approaches. In particular they offer a powerful set of statistical measures for testing alternative model components and checking the fit to data.

8.8.6 Matrix estimation issues

Reviewing matrix estimation methods for this chapter has raised some issues which do not directly affect model estimation.

8.8.6.1 Estimation without a prior matrix or trip end information

Cost deterrence, based on choice theory and travel cost information, provides a strong structure for trip distribution models. Nevertheless, trip end data is still needed to calibrate a model from screenline counts. This emphasises the difficulty of estimating a good matrix from counts alone with only an arbitrary structure determined by the scopes of the counts. This is recognised as bad practice.

8.8.6.2 Non-optimality of multiplicative form

Spiess has shown that the multiplicative form of the log-linear model is not strictly the maximum likelihood solution with Poisson errors. The existence of better solutions outside the log-linear framework may help interpret the behaviour of models constrained to it.

8.8.6.3 Bias and trip splitting

No theoretical work has been found on bias in matrix estimation since Maher's paper in 1987. However, there remains a concern among practitioners that matrix estimation tends to produce more shorter trips. This was one of the concerns that lead to the re-modelling of Wellington with the WTSM.

Irving et al reported the problem in Tyne & Wear. Maher identified the method they had been using as information minimisation, applying an average rather than a product of count parameters. He showed that although there is no bias with this method under uniform growth, there is a bias towards shorter trips with uneven growth in trip ends. However, the problem is believed to occur with the SATURN suite, whose SATME2 matrix estimation program uses maximum entropy, applying the product of count parameters (equation 13.1, SATURN 10.2 user manual). Maximum entropy is recognised to be biased towards longer trips in the presence of overall growth, because all count parameters will tend to reflect this growth and longer trips will be multiplied by more parameters as they pass through more count sites.

A theory suggests that short trips are favoured for their ability to fit between counts and hence are a response to redundancy in constraints, and are perhaps sensitive to inconsistencies between them.

SATURN is founded on equilibrium assignment, which may be run iteratively with its matrix estimation (figure 13.1 and section 13.1.7, SATURN 10.2 user manual). Bias in the estimation may be an artefact of overconstraint in this heuristic approach compared with a bi-level optimisation or because shorter trips are less readily re-assigned.

8.8.6.4 Bias and trip distribution effects

Van Zuylen (1981) and Bell (1983) introduced an overall factor to account for overall growth, and Irving et al (1986) and Maher (1986) extended this to trip end factors, which are incorporated in MVESTM. These were tested and found useful where there had been corresponding changes in the prior matrix, ie overall growth or individual trip end growths. These are changes in trip generation. It may be interesting to consider whether matrix estimation methods are robust to equivalent changes in distribution, which might be:

- main effect: scaling all costs up or down, equivalent to a change in the cost coefficient for an Exponential deterrence function
- second order effects: differential changes in cost or its coefficient between mutually exclusive sets of attributes, eg:
  - road class (affects parts of trips)
  - user class (white and blue collar)
  - mode
  - generalised cost – balance between time and distance
  - Tannerised cost – balance between cost and log(cost)
  - cost band.
- specific effects: changes in costs due to specific interventions such as a new link or service.

As a generality, any fitting to a structure that is not the true model is open to bias. Bias will not appear if it is defined by a measure that is conserved in fitting to the structure, such as trip end totals where trip end parameters are fitted, as in MVESTM after Irving et al (1986) and Maher (1986). Fitting a cost coefficient conserves the total cost of travel (section 3.3); avoiding bias in this key attribute may justify including this element of gravity modelling in empirical matrix estimation. However, this cannot calibrate the full effects of trip distribution in the presence of all the count parameters and a prior matrix.

The only model for which simple matrix estimation is well specified supposes that traffic grows (or shrinks) when you count it.

8.9 Summary of model estimation from aggregate data

Trip distribution can be calibrated from screenline counts, which are cheap compared with the roadside or household interviews usually required. This has been demonstrated with the MVESTM matrix estimation package but there is a history of 'model estimation' from aggregate data.

In the case of Wellington, trip distribution calibration also needs trip end totals for contrast with the screenline counts. Trip end totals appear to be needed in any practical models; land-use data may provide sufficient information about relative trip ends.

The models have been assessed using the statistical tools available in maximum likelihood fitting. These can also be applied to matrix estimation and building but require the accuracy of input data to be known.

Expected accuracies for screenline counts were interpreted from EEM and DMRB validation criteria together with some analysis of local data but levels of confidence in the input data could not be set with any certainty.

Within this limitation, trip distribution effects are highly significant. Counts at four screenlines (one a cordon around the central area) provide about half the information obtainable from the observed matrix of zone-to-zone movements in a 2500 household interview survey. However, there is less power of discrimination between different forms of distribution or justification for more sophisticated ones such as the Tanner function, possibly due to the level of aggregation in the screenline counts.

The analysis has addressed the conversion between PA daily demand and OD hourly assignment matrices by incorporating period and direction factors in the intercept proportions. This method can also be used to fit cost effects in general matrix estimation. Only a single purpose, commuting, has been analysed. The general methodology of fitting a log-linear model by maximum likelihood appears capable of handling multiple purposes, but MVESTM is not formulated to do so and practical screenline counts may offer little discrimination. The general methodology appears capable of combining data at different levels of aggregation, including disaggregation to households, persons or trips. It is not able to fit hierarchies of choice for nested or mixed logit models, limiting its application to joint trip distribution and mode split models.

The synthetic trip distribution gives a good fit to the screenline counts on which it was calibrated in Wellington, close to New Zealand EEM and UK DMRB criteria for model validation. This goodness of fit depends on adjustment to trip ends in accordance with their errors. The validation criteria cannot be met if trip ends are fixed as in the conventional synthesis of a demand model. The criteria can be met easily if there is no constraint on trip ends as is common in matrix estimation. Sector factors can neatly capture the major adjustments to trip ends for the demand model allowing it to replicate the fit to screenlines. The sectors do not need to match the screenlines, but it is best to avoid major exclaves where screenlines depart from natural boundaries.

The method might be used for the final adjustment of a demand model calibrated from a sophisticated analysis or assembled from imported parameters, in the same way that revealed preference data is used to adjust stated preference analyses.

9 Conclusions

9.1 Generalised linear models

This study has shown that the statistical methods of generalised linear models (GLMs) are an effective and flexible method for calibrating trip distribution, in the context of one major working transport model.

Previous work has shown that theories of entropy and random utility can lead to an Exponential model of the form

$$t_{ij} = P_i A_j p_i a_j \exp(-\lambda C_{ij})$$

The cost coefficient λ can be determined from a 'four-square' set of trips and costs between two production zones and two attraction zones.

In a larger matrix, there will be redundancy of information, and error in practical observations. GLMs can find the best fit under maximum likelihood of Poisson sampling errors to a log-linear model that represents the multiplicative form of the trip distribution model. The Exponential deterrence function shown above is the natural form for log-linear GLMs, which have been used extensively for accident analysis.

Trip distributions fitted by this method replicate key measures in the observed data, corresponding with coefficients fitted in the model. In particular:

- Balancing factors p_i and a_j ensure that zonal trip ends, and hence total trips, are replicated.
- The cost coefficient λ replicates total travel cost, and hence (with total trips also replicated) average trip cost.
- Any constant 'K' and cost coefficient 'L' factors for segments of the matrix replicate trips and costs over their respective scopes.

These are inherent properties of a log-linear model with maximum Poisson likelihood. Desirable as they may be, they cannot be used to demonstrate the validity of the model's fit to observations.

A wide variety of other deterrence functions can be fitted by GLMs. These included empirical step-wise deterrence functions which can be fitted by simpler Furness or Fratar iterative balancing methods. However, these step models are shown to be liable to overestimate total travel costs.

Zonal trip matrices were fitted, as is usual in trip distribution modelling, but GLMs also fitted data disaggregated to trip level or intermediate person or household levels. These all gave the same coefficients and measures of significance where there was no loss of information by averaging of factors which varied at lower levels.

This study has been based on a low sampling rate for a household interview survey (HIS). The hypergeometric distribution sometimes replaces the simpler Poisson to reflect high sampling rates from a trip distribution. However, the concepts of a random allocation of trips to a trip distribution under random utility or entropy suggest that a Poisson distribution may still be adequate.

Coefficients from synthesised models can be recovered exactly by GLMs specified with error distributions other than the Poisson, because an exact fit can be achieved. Different error distribution specifications give different results from observations with substantial sampling or other errors.

9.1.1 Measures of fit

GLMs allow the fit of models and their components to be assessed against the uncertainty arising from errors. Their statistical properties follow those of simple linear regression but with a degree of approximation. The accuracy of these approximations can depend upon adequate numbers of observations, and observed trip matrices are liable to be sparse with far fewer original, unexpanded observations than matrix cells.

The principle statistical measure is the deviance, a function of the likelihood which is maximised in GLMs and equivalent to the sum of squares in simple regression. The mean residual deviance of a log-linear model, as applied to trip distribution, is expected to be unity for non-sparse data and this can be used as a test of the model's overall fit. However, this expectation varies and ultimately tends to zero with sparsity.

Analytical approximations for sparse residual deviances have been found. The relative standard error of the mean is shown to depend on the number of trips observed, rather than the number of matrix cells (and hence data records input to the GLM) as would be the case for non-sparse data.

Under sparsity, the expected residual deviance is a function of the fitted model and not of the observations so it varies between models. It has been calculated for many models by elaboration of the deviance over the distribution of Poisson outcomes. The expectation of the residual follows its fitted value quite closely, rendering it uninformative about the fit of any particular model. It appears to reflect the adequacy of the random model, ie the scale of weighting, as much as the systematic modelling of trip distribution.

Although the residual deviance is sensitive to sparsity, the change in deviance between two models is robust and can be used to assess the significance of changes between them.

The t statistic (estimated mean/standard error) is also produced by GLMs. In simple regression it is equal to the square root of the change in deviance when the single term being tested is added to the model. In the GLMs tested, this correspondence generally holds up to the margins of significance, ie $t<\approx 2$. It is less consistent for more significant terms and the t statistic tends to underestimate the significance of a simple cost term when it is first introduced. Larger differences occur in comparing non-linear models. In these cases the change in deviance, related to the likelihood maximised in GLMs, is the preferred measure.

Pearson's chi-squared statistic is also equivalent to the sum of squares in simple regression, but is not optimised under GLMs. It appears wholly unreliable as a test statistic for sparse data.

9.1.2 Weighting

The statistical measures of fit depend on the weighting of observations. Where the principal error arises in sampling, weighting depends on the expansion from the survey to the whole population. In practice, expansion schemes can be complex but GLMs allow a different weight for every item of data.

This can be calculated from the unexpanded observed counts and the expanded trips, but only for matrix cells where a count has been observed. Zero cells, where the movements could be observed but none occurred in the survey sample, also need proper weighting based on sampling rates to reflect the information in the observed count of zero.

It appears sufficient to apply a single weight to the whole of a survey designed with a single sampling rate, even if the final expansion scheme is more complex. A GLM with differential weighting replicates weighted averages of trips and travel costs, which are harder to check and interpret.

Common weights are reduced by uneven sampling and expansion factors within their scopes, for which allowance can be made.

A HIS is liable to sample two or more trips to and from the same workplace by the same worker. These trips are not independent observations and the weighting can be adjusted to treat the worker or workplace as the independent unit of observation.

'Empty' zones, with no observed trip ends, giving a row or column of all zeros in the observed trip matrix, do not contribute to the calibration. They are liable to cause computational problems and dilute residual deviances.

Unobserved matrix cells, for example those movements not intercepted by an interview screenline, should be excluded from the dataset. A model can still be calibrated on remaining 'four-square' sets of data, allowing partial matrix methods. These were not necessary in the study because a fully observed trip matrix from the Wellington Transport Strategy Model (WTSM) HIS was provided by Greater Wellington.

9.1.3 Fit to Wellington data

The WTSM study area is large for an urban or conurbation model, so there is a wide range of costs for internal trips. Smaller study areas may not offer such a range for calibrating a deterrence function and trips to and from external zones may be more important.

Calibration of the sophisticated joint distribution and mode split model developed for the WTSM, with over 30 constants and cost coefficients (K and L factors) segmented by household car availability and a geographic hierarchy, was replicated by GLM. After this, analysis was based on a simpler, single matrix of internal car commuter trips.

The introduction of cost deterrence was always hugely significant. Even crude models such as two steps (short or long trips) or crow-fly distances picked up the majority of the effect, in terms of the reduction in deviance produced by the best-fitting deterrence functions.

The Exponential deterrence function with a simple generalised cost term fitted very well, accounting for some 96% of that achieved by the best fits.

Significant improvement was still possible and a wide variety of functions from step-wise to splines produced curves that were concave with respect to the Exponential plotted as a straight line. This is consistent with cost damping investigated by Daly (2010) and earlier summaries of cost coefficients by Bly et al (2001).

The Tanner function, simply adding a log(cost) term to the Exponential's cost term, accounted for most of this difference in curvature, though better fitting models still tended to be concave compared with the Tanner. The Tanner gave a better fit than the hierarchical geographic segmentation of K and L factors used in the WTSM.

The Power function, comprising just the log(cost) term and akin to the gravity formulation, generally gave a worse fit than the Exponential.

The Power function was very sensitive to the formulation of intrazonal and external costs. By its nature, the Exponential is insensitive to these, and the Tanner appeared relatively insensitive too. The Tanner also replicated interzonal travel well, and fitted better than separate Exponential coefficients for internal and external trips. The Tanner accounted for much, but not all, of a difference found between the internal (household interview) and external (roadside interview) datasets. More realistic external costs did not help to resolve these. Of a variety of formulations for intrazonal costs considered, the WTSM formulation, half the minimum interzonal cost with a cap, fitted as well as any.

Separate components of generalised cost (time and distance from AM and IP assignments) were fitted in a GLM. Apart from showing time to be more significant than distance, the dataset did not provide strong evidence for the composition of generalised cost, or its interpretation as a value of time.

The Genstat package allows non-linear extensions to GLMs. Functions such as the Box-Cox and log-normal improved on the Tanner by relatively small but still statistically substantial amounts. Fitting non-linear models is less robust and more of an art than plain GLMs; only the Box-Tukey converged with more than one non-linear coefficient.

Practical measures of fit compared modelled traffic volumes with screenline counts, and the users and benefits of schemes from observed and fitted matrices. These did not show any clear improvement in fit beyond the Exponential model, and suggested some variation in trip distribution beyond the modelled effects of generalised cost.

9.2 Hierarchical generalised linear models

Hierachical GLMs can incorporate a variety of errors in addition to the Poisson sampling error fitted by GLMs. These extra error structures can be used to represent different levels of choice, as can occur between destination choice and mode choice, and spatial effects in trip distribution beyond those of cost deterrence. However, attempts to fit mixed logit and geospatial models by a hierarchical generalised linear model (HGLM) were unsuccessful.

9.2.1 Mixed logit models

Mixed logit models can fit any form of random utility model, and are at the forefront of choice theory. The theory is advanced in this study that they can be fitted by HGLMs and no contradiction to this theory has been uncovered.

However, HGLMs have failed to recover the coefficients used to generate randomised datasets representing a simple nested logit model of the ‘red bus, blue bus’ dilemma. The original problem appeared to present particular difficulties for HGLMs because:

- The observation of the all-or-nothing choice added a great deal of randomness to the underlying probabilities. This requires a large sample to detect effects of mixing or nesting in the underlying probabilities. HGLMs performed better with the continuous underlying probabilities as the independent Y variables, rather than the discrete outcomes from them.
- In typical disaggregate modelling, each decision is based on a separate draw from the mixture of tastes. HGLMs performed better with larger groups sharing common tastes.
- The mixed model formulation may not be so well conditioned as the equivalent nested logit. The Biogeme algorithm used for comparison failed to fit a mixed logit without good initial values, which were not necessary for nested logit.

An incorrect order of choice modelling (ie the larger cost coefficient λ first) does not necessarily give perverse results.

Joint calibration of trip distribution and mode split with different cost coefficients cannot be achieved by simple GLM and still requires specialist logit programs such as Biogeme or Alogit. The practical alternatives for handling the large number of alternative destinations with logit programs currently appear to be:

- Limit the number of balancing factors, by sector rather than zone, as in PRISM.
- Balance trip ends by Furness iteratively with nested logit, as in TMfS07.
- Accept very long run times, as in Christchurch.

9.2.2 Geospatial models

Spatial correlations were sought in the Wellington car commuter trip distribution, either as variances between K factors at successive levels of segmentation or as higher correlations between movements with less separation.

The double-ended nature of trip distribution, with both productions and attractions in two-dimensional space, needed computational resources of the dimension of the fourth power of the number of zones. This limited calculation to fewer, larger sectors missing the direct effects of short-ranged correlations. No clear evidence of spatial effects beyond those of generalised cost was found.

Analysis was also complicated by the inevitable irregularity of zoning, whose effects were addressed by regularisation, and again by the randomness of all-or-nothing choice outcomes. By comparison, basic geospatial theory models regular or block patterns and normal error distributions from observations in simple two-dimensional space.

No method of aggregating costs as zones are combined into sectors gave the same calibration of the deterrence function at all levels of aggregation. This appears to be impossible in general. However, for Wellington the variation in the cost coefficient was not as severe as for some parameters studied in this multiple areal unit problem.

9.2.3 HGLM development

Although HGLMs offer much potential for advanced investigation and calibration of transport models, they are not yet practical for full-scale application to the mixed logit and geospatial models considered here. They do not yet have the maturity and robustness found in their component GLMs to solve the large and perhaps poorly conditioned problems. There is not the same depth of experience in applying them and interpreting their outputs, in particular the h-statistics whose computation can be more demanding than fitting the model.

9.3 Calibration from aggregate count data by MVESTM

Aggregate counts are usually cheaper and easier to obtain than surveys identifying individual production–attraction movements. There is a long and continuing history of calibrating models from counts. Trip distributions have been calibrated using the MVESTM matrix estimation software. This fits multiplicative models by maximum Poisson likelihood, the same methodology as GLMs. The same statistical tests can be applied to model components if the input data is given realistic weights (or ‘confidences’).

Trip distributions have been calibrated from counts on four screenlines in Wellington. Trip end information is also needed and was taken from the WTSM generation model. The fit of the model to screenline counts depends on the confidence in the trip ends. If the trip ends are fixed, as is usual for model synthesis, the trip distribution cannot fit well at the screenlines; if the trip ends are allowed to vary freely, as is common in matrix estimation, the screenlines fit easily. With approximate confidences in screenline and trip end data, the fit at screenlines meets EEM (NZTA 2008) standards for assignment validation.

The all-day production–attraction trip distribution was calibrated from counts by period and direction (origin to destination) by incorporating period and direction factors in the intercept file. These factors might be fitted as parameters.

Costs and functions of costs can also be presented as intercepts to calibrate their coefficients for deterrence functions, without using the specific parameters (alpha and beta) incorporated in MVESTM.

MVESTM's ability to separate data and model structures gives it great flexibility in fitting models from a variety of data sources, particularly at different levels of aggregation. However, like GLMs, it is unable to deal with different levels of choice.

9.4 Sample sizes

A sample of a thousand households, or trips for purposes with lower trips rates, appears sufficient to distinguish the Tanner model from the Exponential, recognising a cost damping effect. Practical measures suggest there are spatial effects beyond those of cost deterrence, but these have not been quantified.

The counts on four screenlines provided trip distribution information equivalent to about a thousand household interview surveys. However, these do not allow detailed or disaggregate modelling and may not distinguish purpose.

Distinguishing different levels of choice in mixed or nested logit models also appears to require substantial samples on a similar scale.

10 Recommendations

Generalised linear models (GLMs) are commended as an effective and versatile method for calibrating trip distribution models. The theory of GLMs is well established and documented and the algorithms are mature, robust and available in commercial and open-source software. A logarithmic link function corresponds with the multiplicative form of the distribution model and a Poisson error corresponds with sampling error from surveys. Different sampling rates can be represented by different weights. The significance of model components can be tested statistically. Complex models with segmentation, K and L factors and (for certain cases) different modes can be calibrated in a single, simultaneous fitting. They can be fitted directly to fully observed trip matrices (without aggregation to cost bands), or to partial matrices such as those observed by roadside interviews.

It is recommended that model developers should adopt the following practices.

10.1 Deterrence functions

- Take the Exponential deterrence function as the base for calibration. It is supported by theories of choice and maximum entropy and is the simplest form to fit by GLM.
- Test for non-linear or cost damping effects in the deterrence function which depart from the Exponential. This may avoid a need for more complicated geographic segmentation. The Tanner function, $cost^{-\gamma}\exp(-\lambda cost)$, can be fitted within the linear form of a GLM. Other forms such as Box-Cox or log-normal may give a better fit, but require non-linear algorithms that are more complicated and less robust. For an empirical exploration of the form of the deterrence function, splines offer a more plausible form than polynomials, which continue to curve beyond the limits of data, or the flat-topped step form of the classic empirical deterrence function. The last can fitted by the simple iterative balancing methods of Furness or Fratar by including cost bands as a third dimension, but is liable to over-estimate total travel. This can be reduced by avoiding wide cost bands even where the data is sparse.
- Treat Power deterrence functions with caution as they are liable to be sensitive to the specification of zoning and to the formulation of intrazonal and external costs.
- Basic GLMs cannot be used for the simultaneous calibration of mode split and distribution where there are different sensitivities, ie cost coefficients for the different choices.

10.2 Survey, coding and data preparation

- Establish a worker's usual place of work in the person record of a household interview survey.
- Count out-and-back trips by a single person as a single observation for weighting trip distribution data.
- Design weighting schemes as part of survey sampling and expansion schemes. The index keys that determine expansion factors should be preserved in the prepared dataset, as should the keys for forming observed matrices, or other datasets for analysis.
- Apply separate weights for separate surveys with substantially different sampling rates. A single weight may be sufficient for the whole of a survey designed to achieve a constant sampling rate.

- Allow for the inefficiency of mixed sample rates in the weighting, by accumulating the square of the expansion factor along with the factor itself.

$$\text{weight} = 1/\text{variance} = \sum(\text{expansion})/\sum(\text{expansion}^2)$$

- Code realistic costs for external links. Although they are unnecessary for a simple Exponential deterrence function, they are needed for any testing of cost damping.
- Remove or omit empty zones, where there are no observed trip ends, from the observed matrix. Otherwise include and properly weight matrix cells with no observations.

10.3 Model building and testing

- Adopt good statistical model building practice of starting from a simple model and only introduce further factors or segmentation that are shown to improve the fit significantly, step by step.
- Use the change in deviance (−2×loglikelihood) to choose model terms and quantify their significance. The t statistic may give contradictory results for large changes in deviance, the introduction of the cost term, or non-linear models.
- Under sparsity, which is likely in practical cases, do not take the residual deviance as a measure of model fit or for scaling dispersion. Do not use Pearson's chi-square statistic (or the Poisson Index of Dispersion).

10.4 Further research

Detailed suggestions for the continuation of analyses are given in individual chapters.

The MVESTM algorithm can use inexpensive count information, but needs a better understanding of the reliability of its data inputs (traffic counts and synthetic trip ends) for a statistical assessment of its outputs. This would also benefit MVESTM's usual application to matrix estimation, and the building and assessment of transport models in general,

Neither GLMs nor MVESTM can accommodate different levels of choice which can occur with mode choice. The 'contraction mapping' used for joint trip distribution and mode split in Scotland iterates between a logit model for mode split and a Furness process for trip end balancing. A GLM might be substituted for the Furness process to give greater capabilities in trip distribution. Alternatively, a logit model might be substituted for one or more of the GLMs that comprise an HGLM to provide the necessary capabilities in logit modelling.

The hypothesis that HGLMs can represent mixed logit choice has not been confirmed empirically. Its demonstation offers cross-fertilisation between the theories and algorithms. A clear refutation would give a better understanding of the boundaries and limitations of both approaches. Either case would broaden the theoretical basis for transport modelling.

11 References

Bell, MGH (1983) The estimation of an origin-destination matrix from traffic counts. *Transportation Science 10:* 198–217.

Bell, MGH (1984) Log-linear models for the estimation of origin-destination matrices from traffic counts: an approximation. *Proceedings of the 9th International Symposium on Transportation and Traffic Theory, Delft*. pp451–470.

Bell, MGH (1991) The estimation of origin-destination matrices by constrained generalised least squares. *Transportation Research 25B(1):* 13–22.

Bell, MGH and S Grosso (1998) The path flow estimator as a network observer. *Traffic Engineering and Control 39*: 540–550.

Bell, MGH; See H J; Grosso S; Axhausen K. c2006. The PFE: a one-stage network flow estimator for transport planning and traffic management. In draft.

Bierlaire, M (2005) An introduction to BIOGEME Version 1.4. Accessed November 2011. *http://biogeme.epfl.ch/*

Bly, P, P Emmerson, T Van Vuren, A Ash and N Paulley (2001) *User-friendly multi-stage modelling advice Item 9.2: modelling parameters, calibration and validation.* TRL Limited for ITEA Division, DTLR.

Brown, HP (1982) The effect of market segmentation on gravity model performance. *Proceedings of 11th Australian road research board conference, part 6*, University of Melbourne.

Carey M, C Hendrickson and K Siddharthan (1981) A method for direct estimation of origin/destination trip matrices. *Transportation Science 15*, no.1: 32–47.

Cascetta, E (1984) Estimation of trip matrices from traffic counts and survey data: a generalised least squares estimator. *Transportation Research 18B,* no.4/5: 289–99.

Cascetta, E and F Russo (1997) Calibrating aggregate travel demand models with traffic counts: Estimators and statistical performance. *Transportation 24*: 271–293.

CN Course note(s) on matrix building, estimation and validation, delivered by Prof H Kirby and others at Leeds University and dated variously up to 1992. See Kirby et al (1992)

Cochrane, RA (1975) A possible economic basis for the gravity model. *Journal of Transport Economics and Policy 3*: 34-49.

Daly, AJ (1982) Estimating choice models using attraction variables. *Transportation Research B 16,* no.1: 5–15.

Daly, AJ (2010) Cost damping in travel demand models. Report of a study for the Department of Transport TR-717-DFT. RAND Europe with PB and WSP. Accessed January 2012. *www.dft.gov.uk/publications/travel-demand-model-cost-damping*

Daly, AJ and JD Ortuzar (1990) Forecasting and data aggregation: theory and practice. *Traffic Engineering and Control 31*, no.12: 632–643.

Department for Transport (DfT) (2006) TAG unit 3.10.3: Variable demand modelling – key processes. Accessed 10 May 2010. Accessed November 2011. *www.dft.gov.uk/webtag/documents/expert/unit3.10.3.php*

Diggle, PJ and PJ Ribeiro (2007) *Model-based geostatistics.* New York: Springer. 228pp.

Emmerson, P (2008) Using the statistical fitting of trip distribution models to examine the validity of some common modelling assumptions. *Transport demand modelling seminar, European Transport Converence, Leeuwenhorst Conference Centre, Noordwijkerhout*, near Leiden, Netherlands.

Erlander, S and NF Stewart (1990) *The gravity model in transportation analysis – theory and extensions.* Utrecht: VSP.

Evans, SP (1976) Derivation and analysis of some models for combining trip distribution and assignment. *Transportation Research 10*: 3–57.

Fisk, CS (1988) On combining maximum entropy trip matrix estimation with user optimal assignment. *Transportation Research 22B*, no.1: 69–73.

Fisk, CS (1989) Trip matrix estimation from link traffic counts: the congested network case. *Transportation Research 23B*, no.5: 331–336.

Fotheringham, AS, C Brunsdon and M Charlton (2002) *Geographically weighted regression: the analysis of spatially weighted regression.* Chichester: Wiley. 282pp.

Fotheringham, AS and D Wong (1991) The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A 23*: 1025–44.

Friedrich, M, K Nökel, P Mott (2000) Keeping passenger surveys up-to-date: a fuzzy approach. *Transportation Research Records*, no.1735: 35–42.

Galwey, NW (2006) *Introduction to mixed modelling: beyond regression and analysis of variance.* Chichester, Wiley: 366pp.

Gaudry, M (1999) SPQR: the four approaches to origin-destination matrix estimation for consideration by the MYSTIC Research Consortium. Montreal: University of Montreal.

Gunn, HF, HR Kirby, JD Murchland and JC Whittaker (1980) The RHTM trip distribution investigation. In *Proceedings of Seminar P, 8th PTRC Summer Annual Meeting, vol P198, University of Warwick*, July 1980.

Gunn, HF, P Mijjer, K Lindveldt and F Hofman (1997) Estimating base matrices : the combined calibration method. In *Proceedings of Seminar E, 25th PTRC Annual Summer Meeting, vol P414, Brunel University*, September 1997.

Gunn, HF and JC Whittaker (1981) Estimation errors for well-fitting gravity models. *Working paper 149*, Institute for Transport Studies, Leeds University.

Hamerslag, R and BH Immers (1988) Estimation of trip matrices: shortcomings and possibilities for improvement. *Transportation Research Record 1203*.

Hansen, WG (1959) How accessibility shapes land use. *American Institute of Planners Journal 25*: 73–76.

Harris, B (1964) A note on the probability of interaction at a distance. *Journal of Regional Science 5,* no.2: 31–36.

Harris R, R Culley and P Davison (Feb 2006) Matrix census tools software - an essential data source for transport planning in the UK. *Traffic Engineering and Control 47,* no.2: 63–64.

Hamre, TN, A Daly, J Fox and C Rohr (2002) Advanced modelling to overcome data limitations in the Norwegian national transport model. In *Proceedings of Applied Transport Methods Seminar, European Transport Conference, Homerton College, Cambridge*, September 2002.

Heagerty, PJ and SL Zeger (1998) Lorelogram: a regression approach to exploring dependence in longtitudinal categorical responses. *Journal of the American Statistical Association 93*: 150–162.

Herriges, J and C Kling (1995) An empirical investigation of the consistency of nested logit models with utility maximization. *American Journal of Agricultural Economics 77*, no.4: 875–884.

Herriges, J and C Kling (1996) Testing the consistency of nested logit models with utility maximization. *Economics letters 50*, no.1: 33–40.

Highways Agency (1996) Design manual for roads and bridges: volume12 Traffic appraisal of road schemes, section 2 Traffic appraisal advice, part 1 Traffic appraisal in urban areas. London: HMSO. Accessed November 2011. *www.standardsforhighways.co.uk/dmrb/vol12/section2/12s2p1.pdf*

Hogberg, P (1976) Estimation of parameters in models for traffic prediction: a non-linear regression approach. *Transportation Research 10,* no.4: 263–5.

Holm, J. T Jensen, SK Neilsen, A Christiansen, B Johnsen and G Ronby (1976) Calibrating traffic models on census results only. *Traffic Engineering & Control 17,* no.4: 137–40.

Holmberg, K and K Joernsten (1989) Exact methods for gravity trip-distribution models. *Environment and Planning A. 21,* no.1: 1–97.

Irving, J, M Oakley and CF Ramsey (1986) The updating of an O-D matrix: a maximum likelihood approach. *Traffic Engineering & Control 27,* no.9: 442–6.

IVT Heilbronn and Sandman Consultants Limited (1999) *Report on new methodologies for harmonisation, combination and merging techniques for origin-destination matrix estimation*. Deliverable 2 of Project ST-97-SC.2101 funded by the European Commission under the Transport RTD Programme of the 4th Framework.

Jensen, T and SK Nielsen (1973) Calibrating gravity models and estimating its parameters using traffic volume counts. *Proceedings of 5th Universities Transport Study Group Conference*, University College London.

Kamata, A (2002) Procedure to perform item response analysis by hierarchical generalized linear model. *Paper presented at the annual meeting of the Florida Educational Research Association, New Orleans*.

Kirby, HR (1970) Normalizing factors of the gravity model – an interpretation. *Transportation Research 4*: 37–50.

Kirby, HR, JM Clavering and others (dated variously up to 1992) *Course notes on matrix building, estimation and validation*. Leeds University, Institute for Transport Studies.

Law, AM and WD Kelton (1991) *Simulation modelling and analysis*. London: McGraw-Hill.

Lee, B (1999) Calling patterns and and usage of residential toll service under self-selecting tariffs. *Journal of Regulatory Economics 16*: 45–82.

Lee, Y and JA Nelder (1996) Hierarchical generalised linear models (with discussion). *Journal of the Royal Statistical Society B 58*: 619–656.

Lee, Y, JA Nelder and Y Pawitan (2006) *Generalized linear models with random effects: unified analysis via h-likelihood*. Boca Raton: Chapman & Hall/CRC. 396pp

Logie, M (1993) Advances in estimating O-D trip matrices. *Traffic Engineering and Control 34,* no.9: 441–5.

Logie, M and A Hynd (1990) MVESTM matrix estimation. *Traffic Engineering & Control 31,* nos.8&9: 454-459 & 534-537.

Low, DE (1972) A new approach to transportation systems modeling. *Traffic Quarterly 26,* no.3: 391-404.

Maher, MJ (1987) Bias in the estimation of O-D flows from link counts. In Proceedings of 19th Universities Transport Study Group Conference, Session F2, Sheffield University. Also Traffic Engineering and Control 28, no.12: 624-7.

Maher, MJ and I Summersgill (1996) A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis and Prevention 28,* no.3: 281-296.

Maycock, G and RD Hall (1984) Accidents at 4-arm roundabouts. *Transport & Road Research Laboratory LR 1120.*

McCullagh, P and JA Nelder (1989) *Generalized linear models.* 2nd ed. Boca Raton: Chapman & Hall/CRC. 511pp.

McFadden, DL (1978) Modelling the choice of residential location. *Transportation research record 673*: 72-77.

MVA (1998). Distribution and mode split model calibration. *London transportation studies technical note 23 B1.10*.

MVA (2005) Multi-modal model data provision for the Denvil Coombe Practice, on behalf of the UK Department for Transport, Integrated Transport and Economic Appraisal Division. Accessed 10 May 2010 from *http://webarchive.nationalarchives.gov.uk/+/http://www.dft.gov.uk/pgr/economics/rdg/vdmc/mmdp/multimodalmodeldataprovision.pdf*

MVA Consultancy (2009) Land-use and transport integration in Scotland (LATIS):TMfS07 demand model development report for Transport Scotland. Accessed 11 May 2010 from *www.latis.org.uk/services/modelling/library/download_reports/TMfS07_NationalRoadModelDevelopmentReport_29102009.pdf*

New Zealand Transport Agency (2008) *Economic evaluation manual*. Volume 1, amendment no.2. Wellington: Land Transport NZ.

OECD Road Research Group (1974) *Urban traffic models: possibilities for simplification*. Paris: OECD. pp98-102.

OmniTrans International BV (2006) OmniTrans version 4.2.8 help file. OmniTrans4.chm. Accessed November 2011. *www.omnitrans-international.com*

Openshaw, S (1984) The modifiable areal unit problem. Accessed November 2011. *http://qmrg.org.uk/files/2008/11/38-maup-openshaw.pdf*

Ortuzar, J de D and LG Willumsen (1994) *Modelling transport*. 2nd ed. Chichester: John Wiley.

Payne, RW, SA Harding, DA Murray, DM Soutar, DB Baird, AI Glaser, IC Channing, SJ Welham, AR Gilmour, R Thompson, R Webster (2009) *The guide to GenStat release 12, part 2: statistics*. Hemel Hempstead: VSN International.

PTV (1997) *VISUM 10.0 user manual*. Karlsruhe: Planung Transport Verkehr AG.

Putman, SH and SH Chung (1989) Effects of spatial system design on spatial interaction models. 1: the spatial system definition problem. *Environment and Planning A 21*: 27-46.

RAND Europe (2004) *PRISM West Midlands. Tour-based mode destination modelling*. (Task 1 report RED-02061-05; also task 12). Cambridge: RAND Europe.

Robillard, P (1975) Estimating the O-D matrix from observed link volumes. *Transportation Research 9,* no.2/3: 123–8.

Sen, A and TE Smith (1995) *Gravity models of spatial interaction behaviour*. Berlin: Springer. 572pp.

Senior, ML and HCWL William 1977) Model-based transport policy assessment. *Traffic Engineering and Control 18*, nos.8&9: 402–406 and 464–469.

Sinclair Knight Merz and Beca Carter Hollings & Ferner (2003) *Technical notes, Wellington transport strategy model.* Prepared for Greater Wellington Regional Council.

Spiess, H (1987) A maximum likelihood model for estimating origin-destination matrices. *Transportation Research 21B,* no.5: 395–412.

Spiess, H (1990) *A gradient approach for the o-d matrix adjustment problem*. Publication 693, CR. University of Montreal.

Suyuti, R, OZ Tamin and K Satoh (2005) The impact of estimation methods in the accuracy of O-D matrices estimated from traffic counts under equilibrium condition. *Proceedings of the Eastern Asia Society for Transportation Studies 5.*

Tamin, OZ, A Sjafruddin, O Purwanti and K Satoh (2003) Public transport demand estimation by calibrating the combined trip distribution-mode choice (TDMC) model from passenger counts: a case study in Bandung, Indonesia. *Proceedings of the Eastern Asia Society for Transportation Studies 4,* nos.1 and 2.

Tamin, OZ and LG Willumsen (1988) Freight demand model estimation from traffic counts. *Proceedings 16th PTRC Summer Annual Meeting*, University of Bath.

Tamin, OZ and LG Willumsen (1989) Transport demand model estimation from traffic counts. *Transportation 16,* no.1: 3–26.

Tanner, JC (1961) Factors affecting the amount of travel. *Road Research technical paper 51*. London: Her Majesty's Stationery Office.

Tanner, JC (1980) Distribution models for the journey to work *TRRL laboratory report 951*. Crowthorne: Transport and Road Research Laboratory.

TModel Corporation (1999) TModel overview. Accessed 13 April 2006 from *www.tmodel.com* (link lost 2011).

TN Technical note: a series documenting the Wellington Transport Strategy Model. See Sinclair Knight Merz and Beca Carter Hollings & Ferner (2003)

Toner, JP, SD Clark, SM Watson and AS Fowkes (1999) Anything you can do, we can do better: a provocative introduction to a new approach to stated preference design *Proceedings of the 8th World Conference on Transport Research 3*: 107–120.

Train, KE (2003) Discrete choice methods with simulation. Cambridge. Accessed November 2011 from *http://elsa.berkeley.edu/books/choice2.html*

Train, K, D McFadden and M Ben-Akiva (1987) The demand for local telephone service: a fully discrete model of residential calling patterns and service choice. *Rand Journal of Economics 18*: 109–123.

Tribus, M and EC McIrvine (1971) Energy and information. *Scientific American 224,* no.3.

Van Zuylen, HJ (1981) Some improvements in the estimation of an OD matrix from traffic counts. *Proceedings of the 8th International Symposium on Transportation and Traffic Theory, Toronto*. pp656–671.

Van Zuylen, HJ and DM Branston (1982) Consistent link flow estimation from counts. *Transportation Research 16B,* no.6: 473–6.

Van Zuylen, HJ and LG Willumsen (1980) The most likely trip matrix estimated from traffic counts. *Transportation Research 14B,* no.3: 281–93.

Webster, A and MA Oliver (2007) *Geostatistics for environmental scientists*. 2nd ed. Chichester: John Wiley. 330pp.

Whittaker, JC (1979) Paper referenced in Gunn and Whittaker 1981.

Wills, MJ (1986) A flexible gravity-opportunities model for trip distribution. *Transportation Research B 20,* no.2: 89–111.

Williams, HCWL (1977) On the formation of travel demand models and economic evaluation methods of user benefit. *Environment and Planning A 9,* no.3: 285–344.

Williams, NJ and SN Beretvas (2006) DIF identification using HGLM for polytomous items. *Applied Psychological Measurement 30*, no.1: 22–42.

Wilson, AG (1969) The use of entropy maximising models in the theory of trip distribution, modes split and route split. *Journal of Transport Economics and Policy 9,* no.1: 108–126.

Wood, GR (2002) Generalised linear accident models and goodness of fit testing. *Accident Analysis and Prevention 34*: 417–427.

Appendix A: Costs

A.1 WTSM storage locations

There is an iterative distribution – assignment loop. The generalised costs are subject to damping by averaging with previous values, held in matrices mf152 et al.

The last, lagged generalised cost is retained in matrices mf44 and 154 in the databank of the base model and in the file results\matrices\HBWgc.311. This is the set of costs used in this study.

A.2 Minimum costs

The lowest cost is 0.4096 generalised minutes, which is the intrazonal cost for zone 58 around Johnston St and zone 64 on the quayside of central Wellington. Following the WTSM protocol for intrazonal costs, both are calculated as half the interzonal cost from zones 58 to 64.

This is the lowest interzonal cost at 0.8192 generalised minutes, comprising:

50m centroid connector from zone 58 to Featherston/Johnston *(node 7354)*

90m along Johnston Street from Featherston Street to Customhouse Quay *(node 7351)*

100m centroid connector from Johnston/Customhouse to Zone 64

240m total

× 15 cents per km vehicle operating cost

÷13.6 cents per minute value of time

÷1.19 vehicle occupancy

0.222 generalised minutes distance component

0.360 generalised minutes time component: 240m @ 40km/h free flow speed

0.582 generalised minutes.

This leaves 0.2372 generalised minutes for congestion effects, which is little more than the 0.222 minutes calculated as the minimum delay at the Featherston/Customhouse Quay signals.

The cost of the reverse movement from zones 64 to 58 is much higher at 9.4455 generalised minutes. This is in part due to Johnston Street being one way, requiring a more circuitous return route.

More importantly, parking charges are applied to movements attracted to zone 58. $2.75 divided between two trips (out and back) and 1.19 persons/car at 13.6cents/min gives a minimum cost of 8.50 generalised minutes. This applies to all attraction movements except intrazonals. The WTSM formulation for intrazonal costs is to take half the minimum cost of any movement to or from the zone, with a maximum of five generalised minutes (10 for public transport). The intrazonal costs can thus be based on movements attracted to adjacent zones without parking charges, such as zone 64 on the quayside of central Wellington. Zone 64 is unusual in not attracting parking charges as adjacent CBD zones do.

Apart from the intrazonal movements from which parking charges are omitted, the extra parking charges on the attraction trip ends have a null effect on distribution with an Exponential deterrence function. The charges are absorbed in the balancing factors.

In the WTSM formulation of joint distribution and mode split, the parking charges should mainly affect the mode split of commuting trips to the CBD. The formulation allows separate constants and cost coefficients (K and L factors) for public transport trips attracted to Wellington centre. There are also separate constants (K factors) for intrazonals, though not specifically in the CBD. The CBD factors are again aimed at modal split rather than distribution but they may allow compensation for the parking charges. The scope and formulation of the factors are varied.

Only 39 households were surveyed within the parking charge areas, with 19 HBW car trips, none intrazonal. The effects of parking charges are thus likely to be negligible on the distribution model as a whole (demonstrated in section 4.10), but may need attention if such a model is applied to issues of inner-city dwellings and densification, or if non-Exponential deterrence functions such as the Tanner are applied.

Neither zone 58 nor zone 64 have any observed HBW car trip productions, so their intrazonal cells are 'empty', and not involved in calibration. The lowest cost for a non-empty cell, ie with trip ends observed in both production and attraction zones, is 0.6066 generalised minutes, for intrazonal movements in zones 88 and 89 in Tawa. No HBW car trips were observed for these intrazonal movements.

Only movements with observed trips appear in the cost distributions in figures 4.1, 4.9 and 4.21. The lowest cost movement with observed HBW car trips is 0.7479 generalised minutes, intrazonal in zone 25 around Aro St and Central Park. The lowest interzonal cost with an observation is 1.2132 generalised minutes from zone 88 to zone 89 in Tawa.

A.3 Maximum costs

The maximum internal cost is 276.0245 generalised minutes from zone 131 near Otaki to zone 215 beyond Masterton. Trip ends are observed for the production and attraction zones, so the movement is included in calibration, but no trips were observed for the movement itself.

The maximum cost of a movement with an observed HBW car trip is 196.029 generalised minutes from zone 17 (Island Bay) to zone 210 (Masterton). There is only one such observation and the next highest observed costs are around 160 generalised minutes to and from Carterton and Masterton.

External<>internal trips are generally excluded from this analysis. The highest cost is 287.6022 generalised minutes from zone 226, SH1 at the study area boundary on Kapiti Coast to zone 215 beyond Masterton. External<>external movements are set to a cost of 999 in the WTSM and excluded from GLM calibration even if external<>internal movements are included.

Appendix B: Land-use formulation

The household level of aggregation is one at which land-use modelling operates, with a geographic focus on places of residence and employment. Such models may be able to explain spatial patterns beyond the effects of trip cost in transport gravity models. When one was proposed for Auckland, the WTSM household interview survey (HIS) was examined to see how closely it could represent such a formulation.

A distribution of places (home and work) differs from the distribution of trips because:

- not all households have workers
- some households have more than one worker
- workers make different numbers of trips.

In particular, some workers do not go to their workplace on the day of survey, so no workplace is identifiable from the WTSM HIS.

After processing the household data to identify as many workplaces as possible, no workplace was identified for 29% of workers, or 35% of households, 23% of which had no workers.

In the UK, only about 60% of employed people make a journey to work on a given day (Harris et al 2006).

Table B.1 Resident workers' travel

	Count of persons	% of resident workers
Visitor	125	~
Not a worker	3322	~
Not asked	20	1
Absent	72	2
Refused	125	4
Didn't travel	94	3
Didn't travel to work	690	20
Workplace identified	2505	71
Total	**6953**	**100**

Travel refers to weekdays, except that some workplaces were identified from weekend travel.

In WTSM trips but not in land-use model

- Visitors
- Trips to work by non-workers
- Escort and serve-pax trips.

In land-use model but not in WTSM trips

- Absent residents
- Young workers
- Lorry drivers
- Weekend workplaces
- Long-stay employer's business.

B.1 Preparation of a land-use dataset

B.1.1 Visitors and absentees

The WTSM database included trips by visitors staying with the households. There were 125 visitors, of whom 77 were not workers. Of the 48 workers:

- 12 refused to complete travel diaries
- 2 did not travel on the survey day
- 24 did not travel to identifiable workplaces
- 10 had workplaces identified.

Since the workers were not part of the usual household, they were generally excluded when associating households with workplaces.

The person records included residents who were absent. Their details were recorded, but there were no travel diaries. There were 116 residents absent; 10 could not be identified. Of the remainder 34 were not workers and 72 were workers. They were generally coded with 'refusals', with no travel diary or identifiable workplace. They were included in the tabulations of households and workers since they were part of the household related to workplaces.

B.1.2 Identifying workers

Workers were identified as having full-time, part-time or casual employment.

Table B.2 Workers per household

Workers per household	Households	Workers
0	596	0
1	731	731
2	945	1890
3	191	573
4	64	256
5	10	50
6	1	6
Total	**2538**	**3506**

Resident workers including absentees, excluding visitors.

In order to associate a single workplace with each household, workers in each household were ranked progressively by:

- not working at home
- working full time
- age, oldest ranked first
- survey order where ages tied.

Single 'household' workplaces were taken from the highest-ranked worker with an identified workplace. The table below shows some characteristics of the worker ranked first and of the worker at the 'household' workplace in multi-worker households.

Table B.3 'Principal' workers in multi-worker households

Within the household, the worker is/has:	Worker			
	ranked first		at household workplace	
	Count	%	Count	%
No workplace identified		9	106	~
Ranked first	*All*	*100*	944	86
Oldest resident	973	80	780	71
Oldest worker	1035	85	827	75
Highest occupation	682	56	628	57
Highest stated income	932	77	853	77
Highest adjusted income	930	77	802	73
Male	902	74	752	68
Sample size -households	**1211**	**100**	**1105**	**100**

Workers exclude visitors but include absentees.

There were 2775 workers in 1211 multi-worker households, or 2.3 per household.

The oldest resident might not be a worker. The oldest worker might work from home, or not be in full-time employment. This was the case in a substantial minority of households.

The stated income took refusals to answer as the lowest category. These incomes were estimated in the adjusted income.

Males were 51% of workers in multi-worker households; 59% in single worker households and 53% of all workers.

Clearly different rankings would be produced by different criteria, in particular occupation, but there would be a considerable degree of commonality. The identification of a single workplace would be similarly affected.

HBW trip rates to/from the 'household' workplace were similar to those to/from other workplaces, around 1.75 trips/worker/weekday.

B.1.3 Identifying workplaces

There were three main indicators of a usual workplace in trip records:

t*Origin/Dest*Type — Trip end purpose coded in the original survey; included mode change such as bus stop, railway station or car park:

- 4 workplace
- 5 other workplace
- 10 home

t*Origin/Dest*Purpose — Trip end purpose after linking legs to remove mode changes, leaving the activity that generated the trip:

2 work

4 employer's business

12 home

tPurpose Trip purpose, based on purposes at both ends:

1 home-based work

6 non-home-based other (not employer's business)

11 home-based work (escort)

Both zone and meshblock (census unit) were given for every trip end.

Table B.4 Sources of workplace

Source	Number of workplaces	Notes
Weekday HBW trip, t*Origin/Dest*Purpose=2	2193	
Other weekday trip, t*Origin/Dest*Purpose=2	230	mainly NHB
Weekend trips, t*Origin/Dest*Purpose=2	42	
t*Origin/Dest*Type=4	11	some drop-offs?
t*Origin/Dest*Type=5 (employer's business)	29	long stay only
Total	**2505**	

Sources were incremental, ie only for workers whose workplaces had not been found from earlier sources. Only one workplace was identified for each resident worker.

There were 15 work-to-work trips. Twenty-nine workers made trips to more than one workplace in different meshblocks. For six workers the meshblocks were all in the same zone so there would be no difference on the usual scale of modelling, and in several other cases the zones were near to one another, suggesting a minor coding error (eg to different sides of a street). The main workplace was identified from having the greater duration of stay or more HBW trips where durations tied.

Delivery of goods or dropping of passengers were generally excluded. Escort trips for HBW or employer's business were generally excluded, but as this was a trip rather than trip end purpose, it was difficult to tell where the escort element lay in the case of non-home based trips.

About 40% of the sample was asked to complete a weekend diary, 20% on Saturday and 20% on Sunday, so only some of the possible weekend visits to workplaces were included. Six workers not asked to complete weekday diaries gave workplaces in their weekend diaries.

'Employer's business' can be confused with the usual place of work in surveys. This purpose was examined for workers without workplaces otherwise identified. There were 707 trips to or from 296 locations (meshblocks) by 120 persons. This was a far greater activity and diversity of location than for trips coded to the usual workplace, so it was likely that much (and quite possibly all) of it was genuine employer's business. However, locations with longer stays were identified as workplaces; over three hours for the only location visited by a worker, or over five hours if the worker visited several locations on employer's business. Excluding 22 workers who worked at home, this yielded a further 29 workplaces.

B.1.4 Trip frequency

Table B.5 Trip frequency

Trips/worker	Workers		Trips	
0	873	~	0	~
1	672	41.5%	672	24.6%
2	842	51.9%	1684	61.5%
3	53	3.3%	159	5.8%
4	51	3.1%	204	7.5%
5	1	.1%	5	.2%
6	1	.1%	6	.2%
7	1	.1%	7	.3%
Totals	**1621**	**100.0%**	**2737**	**100.0%**

Table B.5 shows the frequency distribution of trips by car between home and workplace by resident workers. Once non-travelling workers were excluded, the average was about 1.7 trips/worker and the major variation was between one and two trips. Since the great majority of tours were complete, starting and ending at home, single HBW trips generally arose from intermediate calls during the reverse direction of travel between home and work.

B.2 Summary of datasets

Table B.6 summarises datasets prepared for land-use formulations, with restrictions to active trip-makers, and compares them with the home-based work (HBW) dataset used in the WTSM and this study, and with the HIS as a whole. Trips are internal to the study area except for the totals for all records.

Table B.6 Size of land use, transportation and household survey datasets

	Zones		Households	Persons	Trips			
	Prod	Attr			Car	PT	Slow	All
Households first workplace	163	189	**1639**	*1639*	1851	414	230	2495
– with any trips	163	180	**1423**	*1423*	*1851*	*414*	*230*	***2495***
– with car trips	157	173	**1086**	*1086*	***1851***	34	18	1903
Persons all workplace	163	207	*1639*	**2494**	2737	620	365	3722
– with any trips	163	196	1491	**2148**	*2737*	*620*	*365*	***3722***
– with car trips	162	192	1218	**1621**	***2737***	58	29	2824
Trips HBW, car	162	194	1224	1740	**3045**	~	~	~
– HBW, all modes	163	199	1497	2267	3045	624	371	**4040**
All purposes & modes	223	224	2397	5495	20,296	1761	4916	26,973
All records, weekday	164	224	2538	6953	20,562	1775	4918	27,898
All records					27,224	1902	6075	35,919

Bold: key values defining row contents

Italics: identical by definition to a value to the left or above

The top two sets of three rows summarise datasets for land-use formulations. In the first set there is one workplace per household and the second set has one workplace per person.

Within these sets, the first line shows all pairings identified between workplaces and households or persons. The second and third lines show pairings with weekday trips, by all modes or by car.

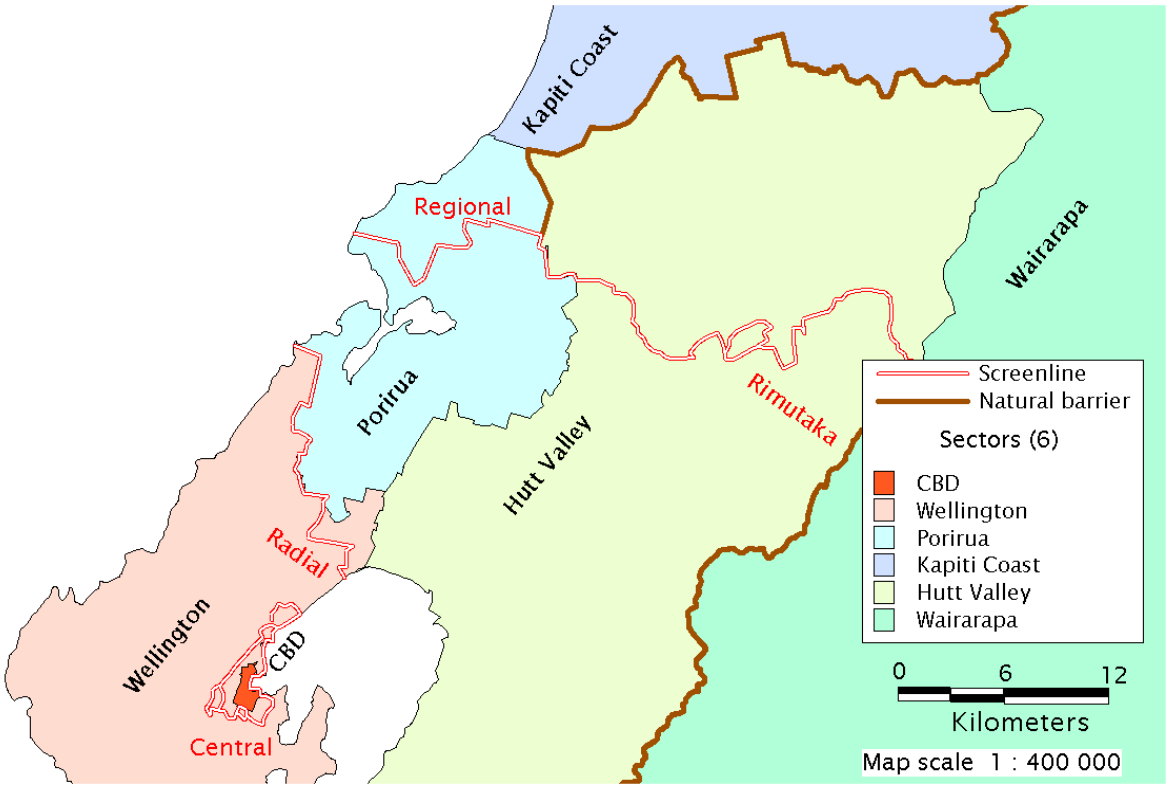
The trips are defined as being included in the HBW dataset used for WTSM demand modelling and in this study, and having the same attraction meshblock as the identified workplace.

For comparison, these 'transportation' datasets are summarised in the next three rows. In the 'all purposes & modes' row, origin zones of non-home-based trips are included in the count of production zones, which elsewhere are just the home zones of households.

The full household survey is summarised in the final two rows. Non-mobile households and persons (without trips) are included in the counts. Trips include externals and their total includes commercial vehicles and other modes not included in the person-trip modelling.

Appendix C: Screenlines and their intercepts

Figure C.1 Screenlines and sectors



C.1 Screenline selection

The screenlines used in the development of the WTSM are described in TN22.1, section 3.1 'Highway assignment validation', where they are identified as L1, P1, P2, P3, U2, UH1, W1, W1A, W4, W5 etc.

Four screenlines were chosen to represent different aspects of a distribution:

1 Central: cordons the city centre of Wellington city

2 Radial: divides inner and outer portions of the region, close to where city-bound traffic is funnelled into a single motorway (SH1/2 at Ngauranga)

3 Regional: cuts the northern SH1 corridor further out

4 Rimutaka: isolates the Wairarapa, and is close to an internal roadside interview site.

Two other screenlines were considered as highlighting other aspects of distribution:

- P2: intercepts radial movements between the Kapiti Coast and the Hutt Valley
- U2 and P3: divide the Hutt Valley and coastal corridors – similar to W5+L1 in intercepting some radial movements into the CBD, but with more 'local' movement.

Other screenlines tended to isolate residential areas, where the imbalance between housing and employment would determine the screenline crossings of HBW trips, largely independent of the distribution modelling.

Screenlines often run on WTSM sector boundaries so they identify intersector traffic. This was not always the case in the chosen screenlines, but there were other criteria for sectors such as consistency with territorial local authorities.

C.2 Derivation of intercepts

The derivation of intercepts can be complicated in practice by multiple and partial crossings of screenlines.

C.2.1 Multiple crossings

If a screenline is complete, without gaps or leaks, any movement between its two sides must be intercepted by the screenline. A 'clean' screenline is only crossed once by any practical route so these are the only crossings. In practice, clean screenlines can be hard to find and some paths cross the screenline more than once; the movements on them contribute to more than one count, and this needs some care in accounting for intercepted movements and count totals.

A particular case is a cordon, which surrounds a small area such as a town centre rather than splitting the whole study area in two. Through trips are a normal feature, often of particular interest; they will cross the cordon twice, inbound and outbound. Again, if the cordon is not clean, other multiple crossings will further complicate the accounting for intercepts.

C.2.2 Partial crossings

Partial crossings arise from multi-route assignment, such as equilibrium which is EMME/2's native method. The WTSM runs up to 140 iterations, giving up to 140 alternative paths between any two points. In practice there are far fewer alternatives but movements are apportioned between them.

This does not affect the total of counts across a clean screenline, because these are all crossings between the two sides of the screenline, each intercepted once and once only. However, where there are multiple crossings, only some of the paths may use them, producing fractions in the intercept matrix.

Cordons are a particular case, where some paths from one side of the cordoned area to another go through the area and others go around it.

Partial crossings are more prominent in leaky screenlines and cordons, where some paths are intercepted by count stations but others go through gaps and are missed.

Unusual paths producing multiple crossings can occur in the first few iterations of an assignment when there are unrealistic loadings and delays, particularly at junctions. These are not repeated as the assignment converges and the associated fractions in the intercept matrix are small.

C.2.3 Method

For a clean screenline, crossing movements can be identified simply from the division of zones lying on either side. Movements between zones on opposite sides of the screenline will cross it, and those between zones on the same side will not.

This was not the happy state of affairs at all screenlines in this study, in particular at the central cordon with its through movements. Some screenlines did not follow zone boundaries, but split zones, and were modelled with centroid connectors on either side of the counted link so the zones were not clearly on one side or the other of the screenline. It was therefore necessary to analyse the routing of all movements to see whether (and how often) they crossed through screenlines. This is called select link analysis.

In EMME/2, select link functionality is provided through the additional options assignment. It can produce only one intercept matrix per assignment run, recalculating paths for each assignment. Trips' AVSELC can generate multiple matrices from multiple screenlines using saved paths, but is limited to seven iterations of multi-routed paths in a file. SATURN has 'pija' ('proportion of I to J assignment') procedures and files associated with its matrix estimation; it saves costs rather than paths.

The WTSM assignment uses particular options for generalised costs and separable assignments of light, heavy and public transport vehicles, and calls extensive subroutines for junction modelling which is not native to EMME/2. The assignment routines have a basic provision for select link analysis, and these have been extended within the WTSM assignment method.

The macros runassau.mac calling assigna1.mac have been modified so that, under the select link assignment option %4 = sl;

to the existing option

%5 active path selection thresholds, eg 0.5,1.5

are added options

%6 path aggregate operator, eg +, .min,.max, traversal or cutoff

%7 additional OD attribute type, specifying intercept matrix contents

1 path

2 active path

3 active path x additional demand

4 additional demand on active path

%8 additional demand matrix, eg mf20, ms1.

The default options are set to '+', 4, and the prime assigned matrix (mf%1) for backward compatibility. The path attribute remains input from link attribute @temp1 and the intercept matrix is still output to mf96. These features are implemented in assignajs.mac, a modification of assigna1.mac.

A macro ScreenPijs takes the name of a screenline as a parameter, which identifies the file of link markers (1 – inbound, 2 – outbound) that are input to @temp2, and names the output files. It prepares scenarios 11, 12 and 13 for AM, IP and PM, and calls assignajs.mac with the appropriate parameters for inbound, outbound and through movements, copying markers from @temp2 to @temp1 as appropriate. Intercept matrices are copied to mf241-9 and output to a text file. Another output file saves the select assignment (volad).

A macro ScreenBat calls ScreenPijs for each of the screenlines in turn, copying the scenarios from 11–13 afterwards.

C.2.4 Output matrices

The prime intercept matrices contain the proportions of paths, rather than the volume of traffic intercepted from a particular matrix. This allows the intercept matrices to simply multiply a new trip distribution matrix to calculate its screenline crossings. In EMME/2 this is a type 1 or 2 additional OD attribute, which is not factored by the assigned or additional demand matrix. This also avoids computational problems where the cells of a demand matrix are small or zero.

Where there are multiple crossings, there needs to be a distinction between the number of paths involved and the number of crossings they make.

The prime intercept matrices contain all paths with any crossings, factored by the number of crossings. This gives consistency between the matrix calculation of screenline crossings and results from an assigned network. This is achieved by marking counted links with a 1 as an additional attribute and then summing the number along each path with a '+' as a path operator. With a type 1 additional OD attribute, all paths are included in the intercept matrix irrespective of any thresholds for active paths.

This allows the thresholds for active paths to be set to identify multiple crossings in the same direction from the assignment of additional volumes to the network. The additional demand can be specified as a unit matrix to show the number of paths, or a traffic matrix to show the volume of traffic involved.

Other specifications are useful for diagnosing the multiple routing paths; in particular, an intercept matrix is generated for 'pure' through traffic, which crosses a cordon once and only once in each direction.

The prime intercept matrices are generated separately for each direction across a screenline. These two directional matrices can simply be added together to give two-way crossings. Both include the movements in the through matrix.

C.3 Output plots

The following plots are defined differently. Inbound and outbound volumes are 'pure' movements that cross the screenline once and once only. The through movements are all other movements, including all multiple crossings in the same direction, though the vast majority are 'pure' through movements, just in and out.

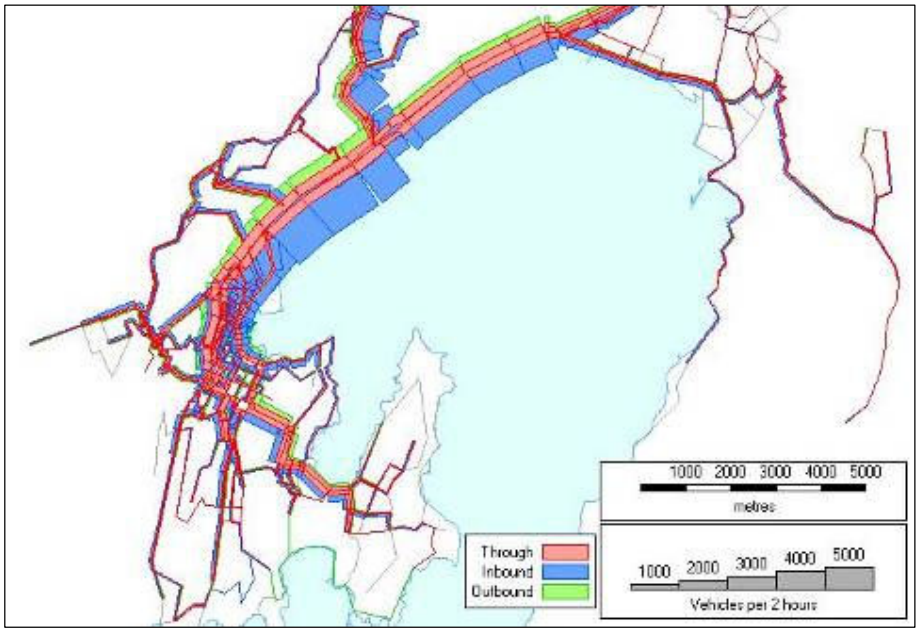
Inbound and outbound are defined as to and from the centre of Wellington city. Plots are shown for the AM period.

The volumes are those of all vehicles used in the final period assignments (mf01-3). They are for two-hour periods, include airport traffic and HCVs but not buses and are factored by matrix adjustment factors.

The paper scale of bandwidths is the same in all the main screenline plots.

C.3.1 Central screenline

Figure C.2 Traffic crossing central cordon – AM



This is a cordon around the centre of Wellington made up from WTSM screenlines W1, W1A and W4. The main screenline W1 is open to the north and does not intercept the main SH1 motorway approach. Screenline W4 completes the cordon but lies some way to the north of the centre. Screenline W1 also passes some way south of the centre and screenline W1A lies closer to the centre, so this has been substituted to give the tightest cordon available around the CBD.

This cordon cuts three zones in the model, where centroid connectors span a counted link:

Zone	Counted link
26	7496–7298
44	7427–7500
73	1590–1589

These do not appear to pose any computational problems, but the model cannot represent local traffic on the counted link exactly.

There is naturally a large volume of through traffic.

The only substantial case of multiple crossings in the same direction is between Kelburn (zone 27) and the north, which passes out of the cordon on Upland Road before returning through the cordon to head north on Glenmore Street. There are very small traces of circuitous routing around Massey University, Central Park and Oriental Parade, probably to avoid overloaded junctions on early iterations.

Figure C.3 Multiple screenline crossings to Kelburn

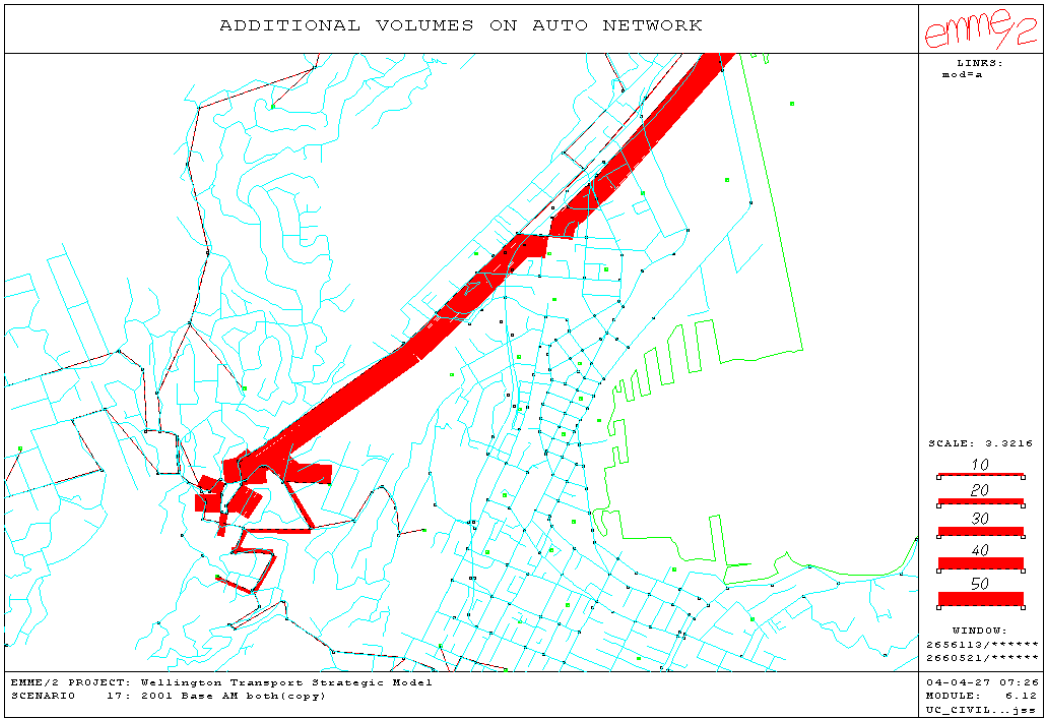
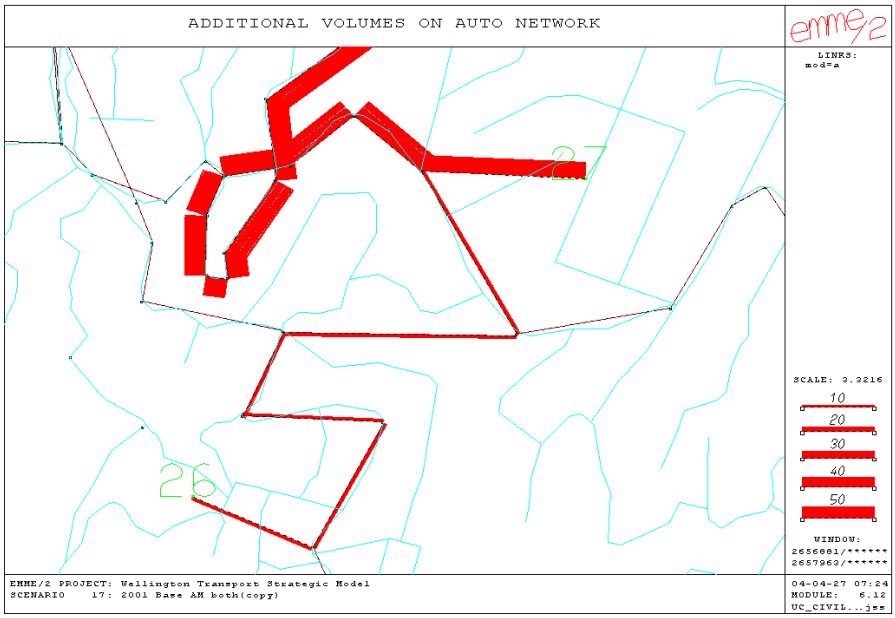
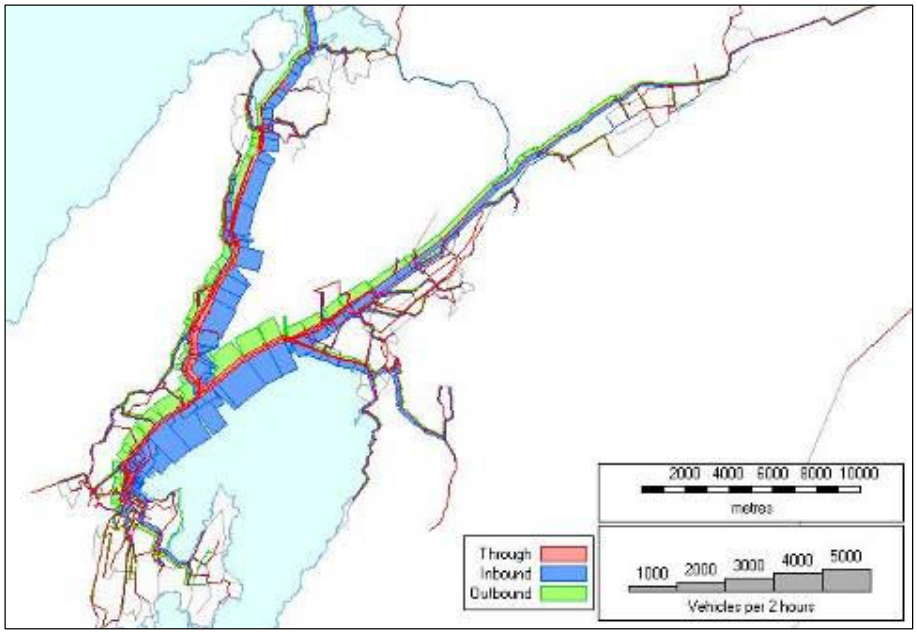


Figure C.4 Multiple screenline crossings to Kelburn – detail



C.3.2 Radial screenline

Figure C.5 Traffic crossing radial screenline – AM



This comprises WTSM screenlines W5 to the north of Johnsonville, and L1 on SH2 Hutt Road north of its divergence from SH1 at Ngauranga. It was chosen to intercept radial traffic approaching Wellington from the north.

However, there are significant through movements, crossing both inbound and outbound. They are principally:

- between the Hutt Valley and Porirua
- between Wellington and Churton Park (zone 82), taking SH1 past Johnsonville to the Glenside interchange (crossing the screenline once) and then going back down Middleton Road to Churton Park, re-crossing the screenline.

The combination of these two movements, between Churton Park and the Hutt Valley produces some multiple crossings of the screenline in the same direction.

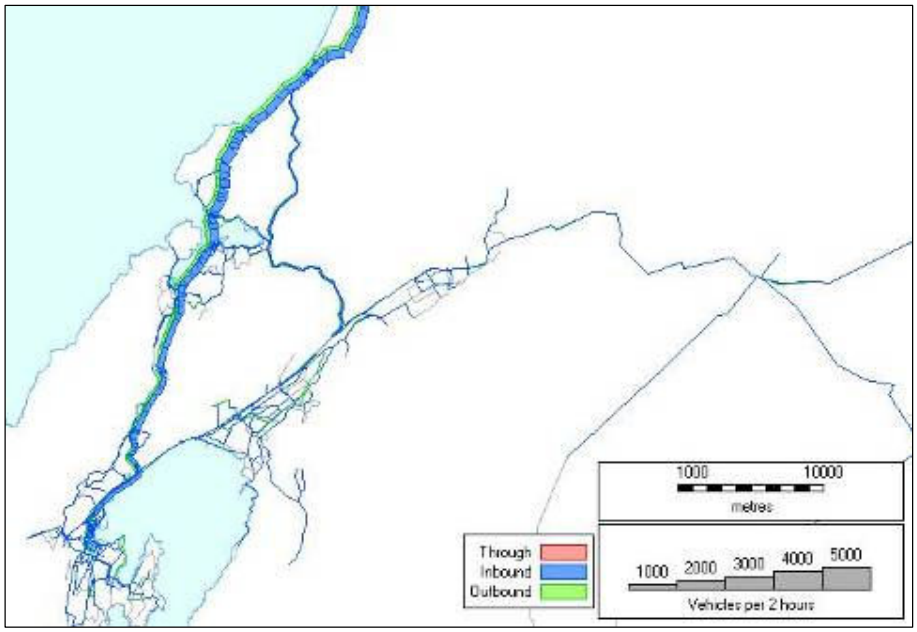
These two through movements also account for most of the partial movements, where only part of the movement is intercepted by the screenline. SH59 provides an alternative route between the Hutt Valley and the Kapiti Coast, and traffic between Churton Park and Wellington can travel via Johnsonville without crossing the screenline.

Other areas of Johnsonville have paths to Wellington via the Glenside interchange that also generate double counts. However only small proportions of the movements have taken these paths; probably a single early iteration of the assignment when there were exceptional delays at a junction.

For a clean interception of radial traffic, screenline W4 may be better, but it is used to complete the central screenline.

C.3.3 Regional screenline

Figure C.6 Traffic crossing regional screenline – AM



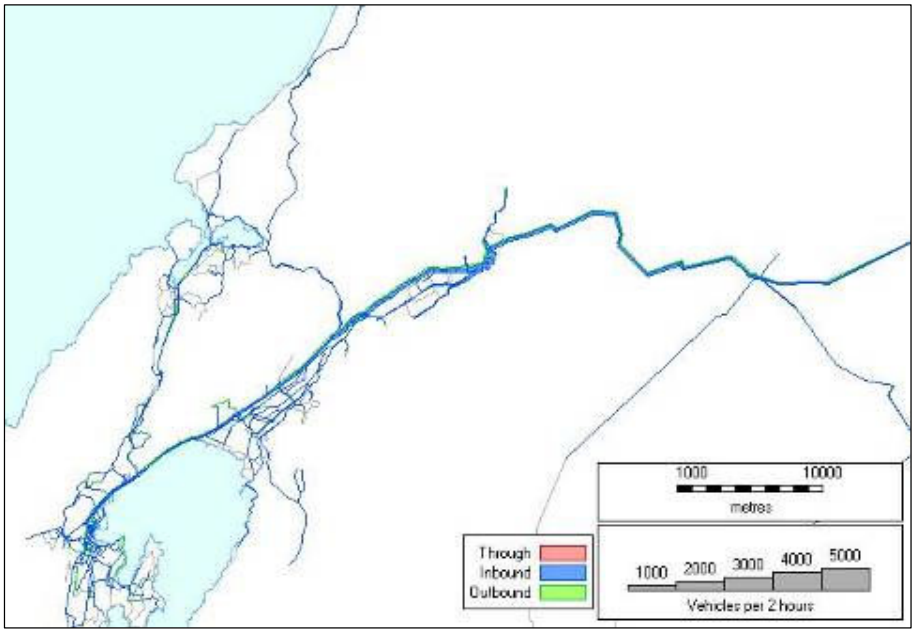
The WTSM screenline P1 separates the Kapiti Coast, served by SH1 running north, from the rest of the study area. There is an area of high ground between the Kapiti Coast and Porirua Harbour which forms a natural barrier.

The screenline cuts both Airlie Road and SH1 south of their junction. This does not generate any through movements using both counted links in the model, and it appears from mapping that few such trips will occur. The count point on Paekakariki Hill Road is described as 'north of Grays Road', so it may include some local traffic as well as that crossing Paekakariki Hill, but the total volume is small.

The intercept pattern is 'clean', without any multiple or partial crossings.

C.3.4 Rimutaka screenline

Figure C.7 Traffic crossing Rimutaka screenline – AM



The Wairarapa in the east of the study area is only connected with the rest of the highway network by SH2 over the Rimutaka Pass. This winding road rises to 555m crossing a barrier of hills that forms a natural screenline, with little development for 15km.

The nearest screenline used in the WTSM's development is UH1, on SH2 south of Akatarawa Road. Akatarawa Road appears to make a tortuous connection with the Kapiti Coast, bypassing the regional screenline as well, but this link is not included in the model.

The link counted for the screenline is 7604–7606. This leaves zones 132, 133, 134 and 136, in the upper reaches of the Hutt Valley, on the Wairarapa side of the screenline. The screenline splits zone 135 both physically, with Maoribank to the west and Timberlea to the east, and in the model, with centroid accesses located at Moeraki Road and Vista Crescent.

However, it does not produce any partial paths because all the zone's traffic is assigned via the shorter Vista Crescent centroid connector to node 7606, so zone 135 appears wholly on the Wairarapa side of the screenline.

The intercept pattern is 'clean', without any multiple or partial crossings.

C.3.4.1 Roadside interviews

A Rimutaka screenline separating the Hutt Valley from the Wairarapa also appears interesting because there was an internal roadside interview survey, giving observations of PA movements.

The interview site locations are SH2 Te Marau; northbound site just south of Te Marau Golf Course, southbound site in old weigh station. These may correspond to links 7600–7601 or 7600–7851. They lie to the east of the UH1 screenline count point, with Akatarawa Road and zones 132, 133 and 136 in between. Zone 134 may also lie in between, or be split by the interview site.

The survey sites are still west of the Rimutaka pass and not all the movements interviewed have trip ends in the Wairarapa on the far side.

There are 2480 interviews (table xScreenline in 2001_HIS.mdb), of which 665 are light vehicles with the purpose of home-based work. The eastern trip end for 935 of all the interviews is to the west of the Rimutaka, rather than in the Wairarapa to the east.

It is likely that a separate count was conducted at the interview site(s) for survey expansion, but this does not appear in the data provided with the WTSM. Interview records of local trips not crossing the Rimutaka could be removed. The movements across the Rimutaka which are missed by the surveys, because they stop short at the upper end of the Hutt Valley, are probably few in number and might safely be ignored. However, because of these complications and since there is no count for a screenline at the natural watershed, it was decided not to pursue this line. No use of this roadside survey is apparent in the development of the WTSM, and there are signs that the data requires editing.

Appendix D: Schemes

Example schemes were set up to show the effect of different trip distributions on the outcomes of models. Unlike screenline crossings, the results did not show whether a trip distribution was a better or worse model, but indicated the sensitivity of model outcomes to the choice of trip distribution.

The schemes were developed from the 2001 base network and loaded with 2001 traffic demands for consistency between observed and calibrated distributions. In practice, a feasibility study would be based on future year models.

D.1 Measures

D.1.1 Benefits

The scheme benefits were calculated from the difference of cost matrices with and without the scheme. The costs were those used for the assignment, measured in generalised minutes. Unlike the costs used for distribution, vehicle operating costs were set at a perceived value of 7.5c per km for fuel including GST. There were no parking charges since these did not affect route choice once the destination had been determined in the distribution stage.

D.1.2 Users

Scheme users were identified by selecting paths through designated links in the scheme, in the same way that screenline interceptions were found. The result was an indicator matrix showing the origin-destination (OD) pairs that used the scheme. Values were in the range 0–1; usage of multiple scheme links was not counted, but partial usage of the scheme arising from multi-routing was represented by a fraction.

Most schemes involved several links forming at least two distinct movements, in two directions. Many definitions of users were possible but for simplicity one key link was defined in each direction. It would be possible, but more complex, to identify those who used the 'whole' of the scheme. The scheme benefits provided a systematic appraisal of all those affected, including those who did not travel on any new link, and allowed for the varying degree of benefit to different users. The definition of users was therefore kept simple.

D.1.3 Relief

It was thought that, as well as the use of the new link, the relief of existing sensitive links might be a useful measure. However, in practice it appeared that the major relief was well reflected by the number of scheme users, and secondary effects were dispersed and not critical to the scheme. It was not worth generating yet more measures of relief to existing links that were either closely correlated to the scheme users, or relatively unimportant to the scheme.

D.2 Methods

The procedures were based on a Wellington Transport Strategy Model (WTSM) procedure for updating networks (2011update.mac), but without any re-initialisation of the public transport model (boardings, fares, centroid connectors). Once the 2001 base network was updated, scheme links defining users were tagged and the 2001 base AM matrix was assigned. The final generalised costs were skimmed to one matrix and movements using the selected links were indicated in another intercept matrix. The base network cost matrix (mf21) was subtracted from the scheme cost matrix to give the change in costs.

The resulting matrices represented scheme effects. With factors for period and direction, they could multiply a trip distribution matrix to estimate the scheme benefits to traffic in that matrix and the volume of traffic in the matrix that used the scheme.

This methodology allowed first order estimates to be made directly from calibrated or synthesised trip distributions without re-running the model. Its simplicity and consistency made it easier to analyse and comprehend.

Variable matrix effects of the scheme were not considered; only differences between the base and scheme networks affected the scheme benefits. There was no feedback of changed costs due to the scheme into the trip distribution stage. These were second-order effects.

The models were run for the morning peak, when home-based work (HBW) trips were most important.

Each scheme was run with a different macro; the three schemes could be run together with a batch macro calling each in turn and storing the scenario afterwards.

D.3 Choice of scheme

Three schemes were chosen:

- a central scheme, in Wellington city
- a radial scheme, on the main approach to the city
- a regional scheme, at the edge of the conurbation.

They offered a variety of circumstances. Such schemes had been mooted at some time and could be modelled for a feasibility study, but none of the codings used here necessarily represented the desires or intentions of any authority.

D.4 Central scheme

The central scheme represented an improvement to the route of SH1 around the centre of Wellington. It extended from the existing motorway where it emerged from its tunnel under The Terrace. A new link allowed the one-way pair of streets carrying SH1 to be shifted by one block further from the city centre.

D.4.1 Specification

The coding was taken from files provided with the WTSM model to update the 2011 network (directory Network\2011Base\futproj\). The updates worked with the 2001 network, provided some bus lines were deleted before the highway network was altered (new file ICBPamptdel.221). There were updates to:

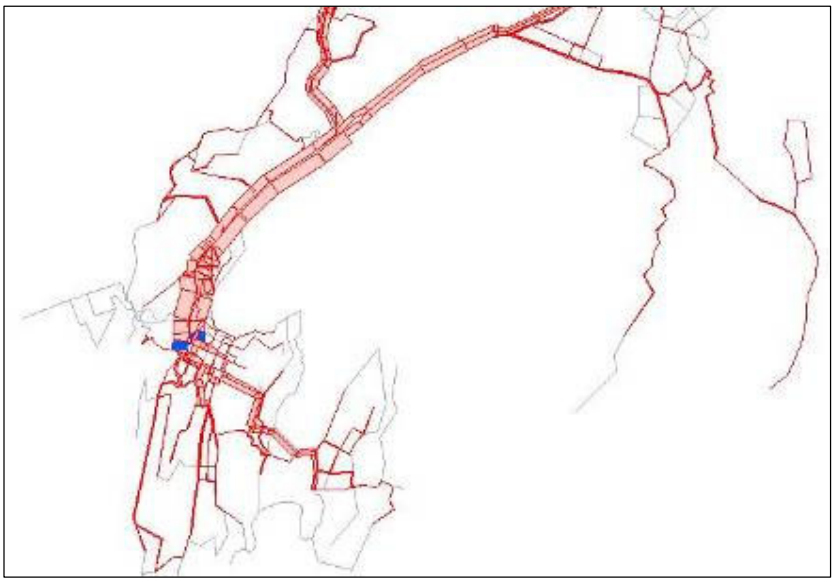
- the highway network structure (ICBP.211)
- nodes requiring expansion for turn reporting (ICBP_turn.231)
- link characteristics (ICBPlxtra.in)
- node characteristics, including junction modelling (ICBPnxtra.in)
- public transport lines (ICBP_ampt.221).

Scheme users were defined as those traversing the new ends to the motorway, immediately west of Willis Street.

D.4.2 Performance

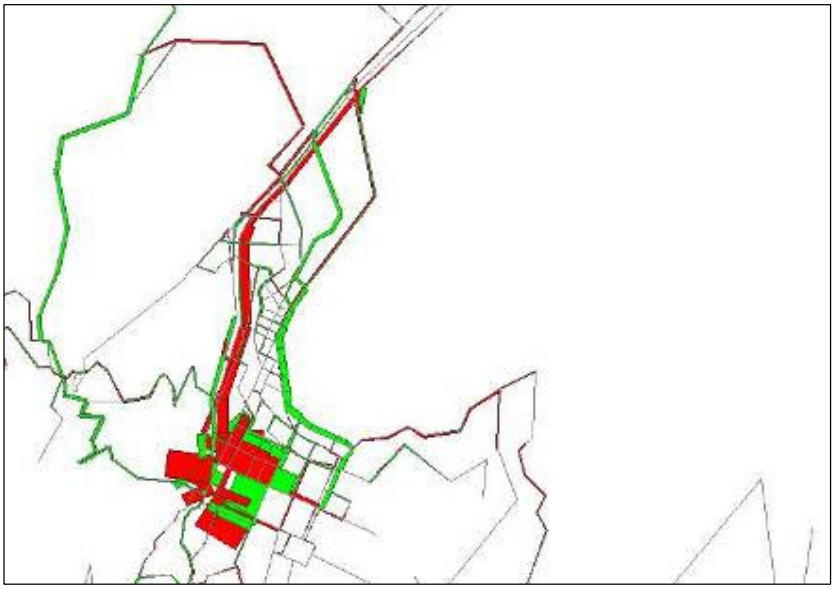
The main effect was to move traffic from the existing designated through routes, Ghuznee and Vivian Streets, to the new routes, Vivian Street and the new link. Although changes were apparent in other parts of the city, they were of lesser magnitude and fairly well dispersed. The quays benefited from one of the larger secondary reductions, but this was a relatively insensitive high-volume traffic route in terms of relief. Although some movements gained benefits of over three generalised minutes, most of the benefits were of two minutes and below, with small disbenefits for a lesser number of movements. This spread of benefits was probably due to adjustment within a congested network.

Figure D.1 Central scheme users



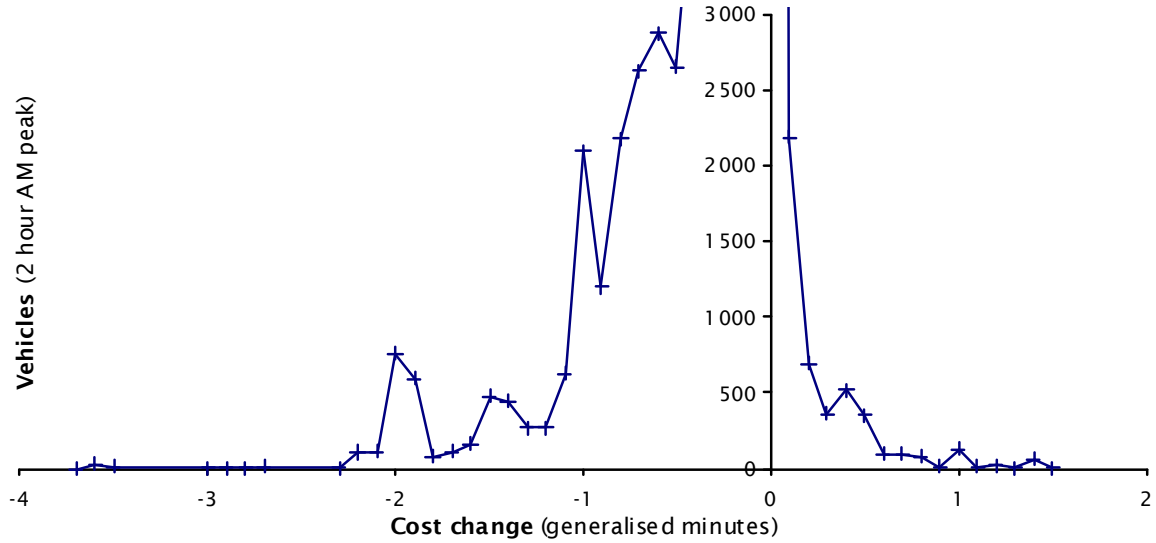
Users of the links shown in solid blue.

Figure D.2 Central scheme changes in traffic flows



Red – increase; green – decrease

Figure D.3 Central scheme benefits



D.5 Radial scheme

The radial scheme represented a widening of the SH1 motorway between the merge of SH1 and SH2 at Ngauranga and Aotea Quay on the outskirts of Wellington. At Ngauranga, two lanes from each of SH1 and SH2 merge into a three-lane section. There is usually congestion at the merge for about an hour in the morning peak.

D.5.1 Specification

The merge was not modelled by a bottleneck, as suggested for SATURN (9.2 App Q) and possibly implemented in earlier versions of the WTSM. All links were coded with the volume delay function for links without a controlling junction, vdf11. There was some variation in the coding of links that might be intended to replicate delays at the merge.

The most notable variation was in Ja .The effect of this parameter appeared to be to increase delays as capacity was approached, akin to the randomness element in the Transyt delay model. It varied between 0.1 and 1.8 on different parts of the motorway and freeway system. It was generally higher than the 0.4 adopted for motorways in table 4.2 of TN14.1, or the 0.1 suggested for freeways by Akcelik in table 4.1. A value of 0.8 (the same as that given in table 4.2 for expressways) was adopted to represent an improvement over current conditions, while still falling short of the best rural motorway conditions.

The link immediately downstream of the SH1/SH2 merge was upgraded from 1800 to 1900 veh/hr/lane, consistent with other motorway sections, reflecting relief from merging problems. (The all-purpose section of SH2 beyond the merge has a capacity of 2000 veh/hr/lane.)

Beyond the northern end of the hypothetic widening, the approaching SH1 freeway down Ngauranga Gorge was coded with high values of Ja. On the presumption that these were intended to reflect queueing from the merge that would be alleviated by the scheme, Ja was recoded to 0.8. The geometric difficulties in the gorge were still represented by lower free-flow speeds of 85km/h, capacities of 1800 veh/hr/lane and 2.5 effective lanes.

Immediately beyond the city end of the scheme, the motorway was coded with speeds of 95km/h and capacities of 1900 veh/hr. The outbound carriageway approaching the Aotea Quay merge had Ja=0.8; the inbound carriageway from the diverge had varied values, but it was not clear how the scheme would influence these. Therefore no changes were made to the links beyond the city end of the scheme.

No changes were made to public transport services.

D.5.2 Performance

The main effect of the scheme was to shift inbound traffic back onto the motorway from the parallel Hutt Road. There was little diversion from the more distant Burma Road route between Johnsonville and Wellington as a whole; the relief on some of this route seemed to arise from the Hutt Road's increased attractiveness for traffic from Khandallah in to Wellington. The benefits of the scheme were typically only 2.6 generalised minutes from the Hutt Valley (SH2), or 1.8 minutes from SH1, perhaps because the base model did not model the full delays at the merge.

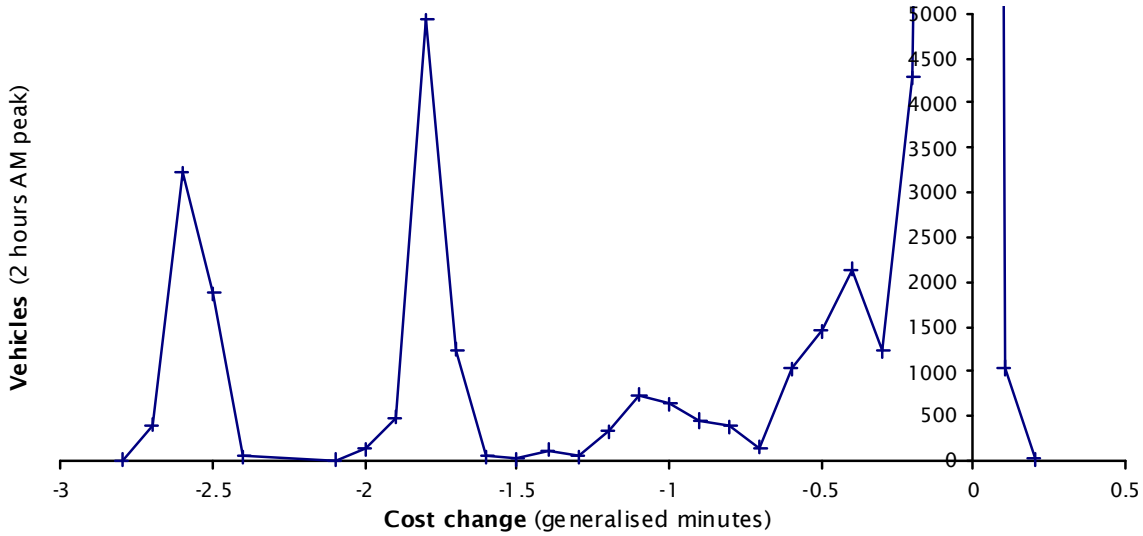
Figure D.4 Radial scheme users

Users of the links shown in solid blue.

Figure D.5 Radial scheme changes in traffic flows

Red – increase; green – decrease

Figure D.6 Radial scheme benefits



Vertical axis: number of vehicles, 2 hours AM

Change in costs (generalised minutes)

D.6 Regional scheme

The regional scheme represented a shortening of SH1which runs north towards Auckland. It avoided a congested bridge at Paremata where SH1 ends as a dual carriageway and improved accessibility to Kapiti Coast within the model area. However, it traverses difficult terrain.

D.6.1 Specification

The scheme was coded very simply as two pairs of one-way links, split by a junction with SH58 that runs between SH1 and the Hutt Valley. At the northern end, the scheme left the existing SH1 between McKay's Crossing and Paekakariki (node1028); it did not bypass McKay's Crossing. It met SH58 east of Pauatahanui, at its junction with Belmont Road (node 7841). At the southern end, it rejoined the existing SH1 freeway between Porirua and Grenada North (nodes 7409 and 7421). The new links of the scheme joined separate carriageways of the freeway so that scheme traffic could only travel to and from the south (Wellington). There were no links from the scheme into Tawa or Ascot Park, so the coding represented a strategic function.

Link lengths were allowed to default to the direct crow-fly distance, which would underestimate the difficulties of the terrain. To compensate, the free-flow speed was set to 85km/h, relatively low for an unrestricted rural link (type 15, usually 100km/h). The capacity was also set low to 1200pcu/hr/lane to reflect the terrain. Only one nominal lane was coded in each direction, but crawler lanes might well be needed to achieve this effective capacity in practice. This seemed sufficient for 2001 AM modelled traffic.

All nodes already existed on the network. No junctions were modelled, for simplicity and because they should impose little impedance in a new design. Link speed-flow was applied with the standard parameter for the road type (vdf 11, Ja 1.4).

No changes were made to public transport services.

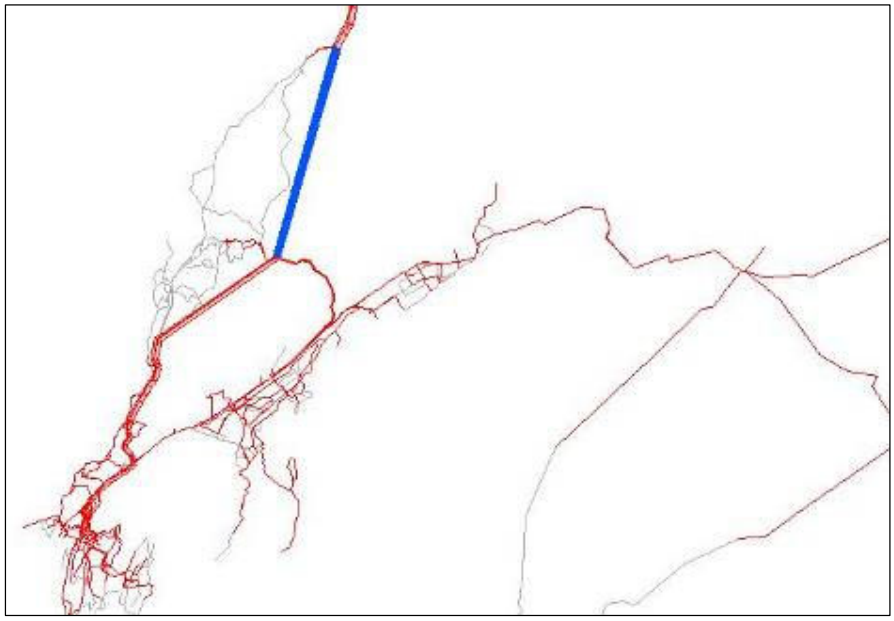
Users were defined as those traversing the northern section of the scheme.

D.6.2 Performance

The main effect was to move traffic from the existing SH1 to the scheme. There appeared to be secondary effects of re-routing movements between Johnsonville and the Hutt Valley onto the southern section of the scheme and SH58, relieving Ngauranga Gorge and SH2. This decongestion produced minor changes towards Wellington as well as within the Hutt Valley.

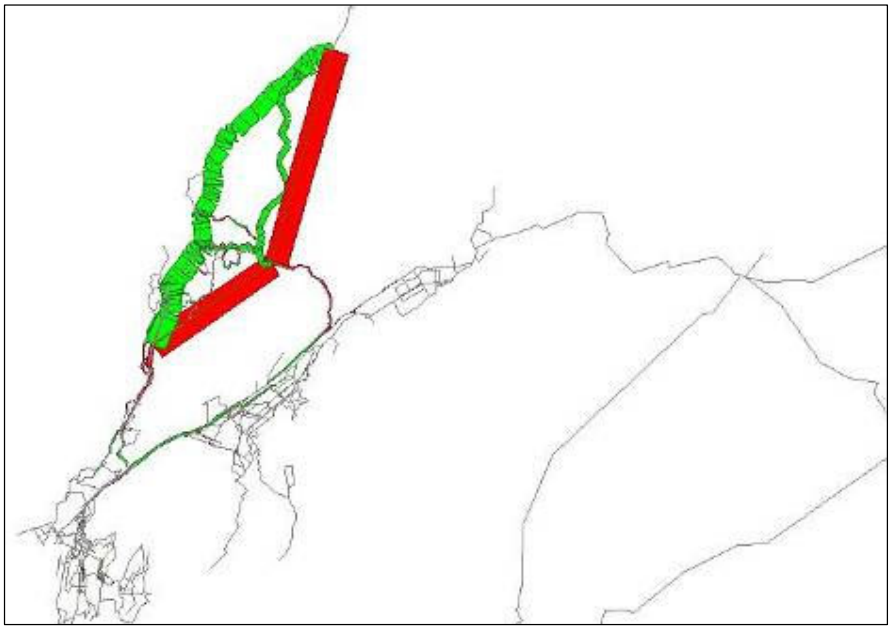
The movements gaining the greatest benefit, almost 10 minutes, were those between the Kapiti Coast and the Hutt Valley, which currently use the slower Paekakariki Hill Road. The benefits for movements using the whole of the scheme from the Kapiti Coast to Wellington were between seven and eight generalised minutes.

Figure D.7 Regional scheme users



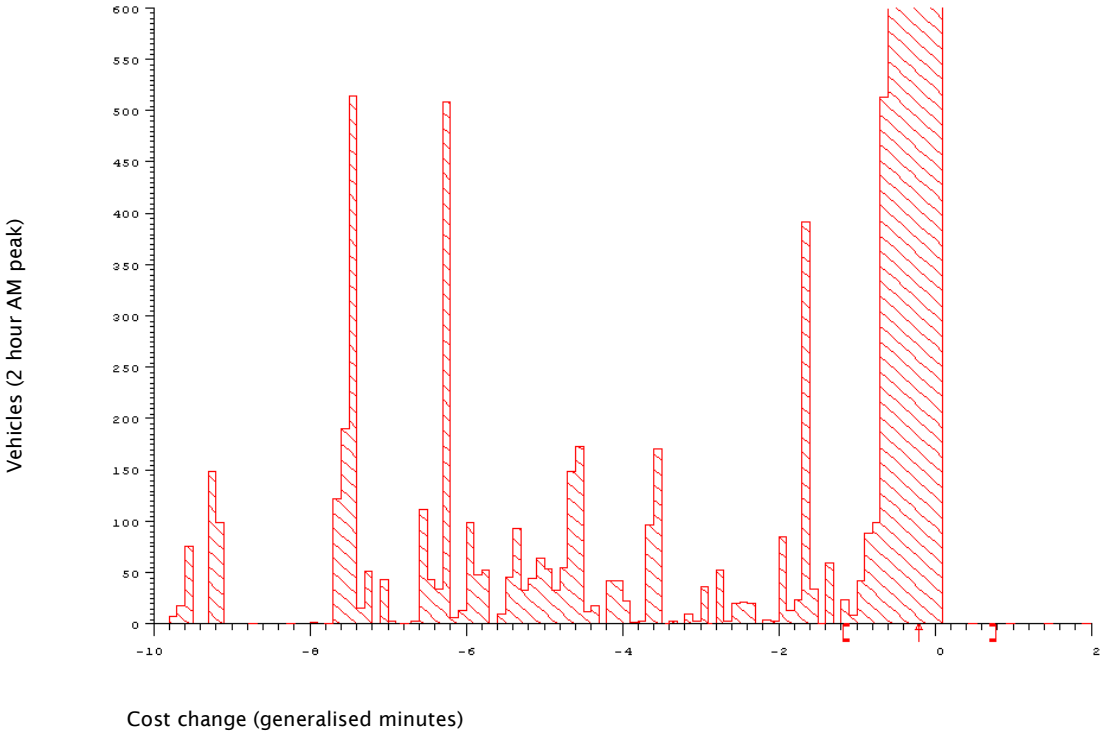
Users of the links shown in solid blue.

Figure D.8 Regional scheme changes in traffic flows



Red – increase; green – decrease

Figure D.9 Regional scheme benefits



Appendix E: MVESTM inputs and coding

E.1 Input files

The following sections elaborate on features of the main input files, in particular their:

- role within MVESTM in the context of table E.1
- features of note
- usage in this study.

They supplement and do not supplant the formal documentation of Trips MVESTM or Cube Analyst.

Table E.1 Juxtaposition of data and model structure in MVESTM input files

Reproduced from table 8.2 in the main text for convenient reference

Information					**Structure**	
Index		Data	Conf	Scope	Index	Parameter
Prior matrix						
Orig	Dest				Orig	Dest	
1 : i : N_{zone}	1 : j : N_{zone}	N_{ij} but not c_{ij}	w_{ij}	*Cell i,j*	1 : i : N_{zone}	1 : j : N_{zone}	fixed as $t_{ij} = N_{ij}$ or $t_{ij} = c_{ij}^{-\gamma}\exp(-\lambda c_{ij})$
Trip ends			*Implicit*	**Model parameters**	
1 : origin i : N_{zone}	O_i	w_i	*Row i*	1 : i : N_{zone}	a(1) [default =1] : a(i) : a(N_{zone})
1 : destination j : N_{zone}	D_j	w_j	*Column j*	N_{zone}+1 : N_{zone}+j : 2×N_{zone}	b(1) : b(j) : b(N_{zone})
Screenlines			**Intercepts**	~	~
Lowest s : screen s : Highest s	Q_k	w_k	Movements ij passing through screen k Proportions R_{ijk}	2×N_{zone}+1 : 2×N_{zone}+k : 2×N_{zone}+N_{screen}	X_1 : X_k : $X_{Nscreen}$
				2×N_{zone}+N_{screen}+1 2×N_{zone}+N_{screen}+2	$-\gamma$ (alpha) λ (beta)

For calibrating a trip distribution from counts, the main files input to MVESTM are:

Prior matrix:	costs
Trip end:	trip end totals and their confidences;
Screenline:	the count, and its confidence, for each screenline;
Intercept:	the set of zone-to-zone movements, or matrix cells, which comprise each screenline count
Parameter:	the starting value, limits and scaling of a parameter applied to each trip end, or to the set of movements intercepted by each screenline, and one or two cost parameters.

These files are bounded by double lines in table E.1.

Abstracts from text files are shown below in fixed Courier font, thus:

```
* Example of a Trips text input file
*   with two lines of data at the bottom
* Next three comment lines just show the format:
*    Anode     Bnode      Flow
*        1         2         3
*2345678901234567890123456789012345
         1         2       800
         2         1      1200
```

Comments are included in the Trips style on lines beginning with an asterisk, which are not read as data.

A full set of input files is given for Spiess' example in appendix I.

E.1.1 Control file

Not shown in table E.1

As for other programmes in the TRIPS suite, MVESTM is usually controlled by settings in a control file. This is a text file, with sections specifying input and output files, parameters (for the program, not the estimated matrix structure), options and selections (not used by MVESTM). Interactive control from the computer console is also possible.

The example below has typical settings for calibrating a trip distribution, some of which are discussed in the following sections.

```
ME calibration
&FILES
IMAT1='CostMatrix.MAT',
IDAT1='TripEnd.DAT',
IDAT2='InputParameter.prm',
IDAT4='Screenline.scn',
IDAT5='Intercept.ICP',
IDAF1='Dummy.RCP',
OPRN='Printfile.PRN',
OMAT1='EstimatedMatrix.MAT',
ODAT1='OutputParameter.DAT',
ODAT2='GradientSearch.GDS',
ODAT3='OptimisationLog.DAT',
 &END
 &PARAM
TABLES=101,102,
```

```
WIDEND=2,
DEC='D',
MFORM=2,
 &END
 &OPTION
TRIPM=F,
COSTM=T,
SCRFIL=T,
TRPEND=T,
INTCPT=T,
MODPAR=T,
 &END
```

Only settings relevant to this study are shown. Other default settings are omitted from this listing and can be omitted from the control file.

This study has used Cube's Application Manager, which provides a window interface for the settings and can automatically name files. The Scenario Manager can insert values of its keys into file names and directories to run a whole catalogue of scenarios with different inputs.

E.1.2 Input matrix file

The prior matrices span the top of table E.1 as they can contribute to both data and structure. As data, the estimated matrix will try to match cells with observed trips. These can be ignored as data by setting their confidences w_{ij} to zero; cost values in cells are never included directly in the objective function.

Both trips input directly and trips synthesised from costs always appear in the structure of the estimated matrix. They are the base values to which other factors are applied. Unlike other factors, their values are always fixed; they are not parameters that are optimised to give the best fit.

Where both trips and costs are input (`TRIPM=T` and `COSTM=T`), the trips N_{ij} are used for the initial cell value t_{ij} if the prior trips are greater than zero; otherwise the cost deterrence functions are used for t_{ij}.

If a cost is being used and its value is zero, the initial cell value is always zero, even for the negative Exponential (alpha=0, beta +ve) or inverse Power (alpha +ve, beta=0) forms of cost deterrence where the function is non-zero for a zero cost.

If the confidence of a cost is zero, the cost (rather than its deterrence function) will be interpreted as a trip, and taken as the initial cell value t_{ij}. MVESTM provides warnings of non-zero trips or costs with zero confidences.

For this study, a set of costs was input via the matrix file, `COSTM=T`. No observed trip matrix was included in the matrix file, `TRIPM=F`.

The costs were been taken from the WTSM EMME/2 model. Their confidences were set at 100 for all cells except those for empty zones, which were set to zero together with their costs. The confidences have no effect on the objective function, but do appear to affect convergence, which deteriorated when the general weighting factor of 100/157.9 was input as the confidence.

The cost matrix and their confidence matrix were stacked in a single input file, `TABLES=101,102`. They were compiled in a Trips format by the MVMODL matrix manipulation program, which also read the trip end file to identify empty zones, and set the corresponding rows and columns of both cost and confidence matrices to zero.

E.1.2.1 Matrix file formats and precisions

Initial runs used the Trips matrix format with its default precision of unity. This was sufficient for most practical purposes, but posed problems when computing deviances from fitted values stored as zero. True zeros should only occur in fitted trip distributions for empty zones, which have no observed trip ends. Zeros from rounding of small values need not present a problem if the corresponding observation is also zero; the resulting deviance is zero, and the true Poisson deviance is also small and may be negligible. The computational problem of taking the logarithm of zero only arises where there are non-zero observations for an expected mean of zero. If the matrix format offers a good precision, only the smallest fitted values are rounded down to zero and it is rare to have observed trips for such movements. However, when estimating from synthetic data, the synthesised trip distributions should have non-zero 'observations' for all cells except those for empty zones, and problems of precision in the estimated matrix become more frequent when calculating deviances.

The Trips format matrix's single precision offers a range of 2.15×10^{9} (4 bytes, or 32 bits, less one bit for sign). This can be scaled up or down by powers of 10, with the DEC parameter in Cube Analyst. The basic trip distribution fitted with an Exponential deterrence has cell values ranging from 2×10^{-6} up to 2×10^{3}. The Tanner and Power functions give smaller ranges, in particular higher minima around 10^{-5} and 10^{-3} respectively, and slightly higher maxima. It is difficult to store all these distributions in a TRIPS format matrix without either truncating large values or rounding small ones down to zero, so the Voyager format matrix with double precision was specified with parameters MFORM=2 and DEC=D in Cube Analyst.

Even when writing its estimated matrix out in the Voyager double precision format, MVESTM appears to have a minimum non-zero cell size of 10^{-8}. This is sufficient to prevent rounding down to zero for all but a few cells in ill-conditioned matrices, eg solutions that are indeterminate for lack of trip end information.

For output to the dBase.dbf format from Voyager, a format of F14.8 was adopted.

E.1.3 Trip end file

Unlike the input matrix file, a trip end file is optional; its inclusion is specified by `TRPEND=T` in the control file. Shown on the left-hand side of Table A, it introduces data which the matrix estimation process will seek to match. In general, the estimated matrix is not constrained to match these trip end totals, as in a simple Furness process; the fit will depend on the structure of the matrix and on the other information it is seeking to match. Trip end totals are not necessarily included in traditional matrix estimations, since they are not readily counted for most zones, but they offer a convenient way to introduce land-use information such as trip end growth.

A similar effect to entering a trip end total can be obtained by specifying a screenline count across a zone's centroid connectors. However, in a TRIPS model these would miss the zone's intrazonal trips, which never appear on the network, whereas the trip end totals include intrazonal trips. In MVESTM, the movements contributing to a trip end total (ie the matrix row or column) do not have to be specified in the intercept file, as they would for a screenline.

The trip end file is a text file, a standard Trips format with multiple productions omitted and added columns for confidences. A sample is given below:

```
* WTSM HBW internal car trip ends from HIS; ordinary weighting
*- Zone ->><---------- Trip End Totals ----------><--- Confidences--->
*         <- Prod ->                    <- Attr -><- Prod -><- Attr ->
*        1         2         3         4         5         6         7
*234567890123456789012345678901234567890123456789012345678901234567890
         1  919.3153                      241.0370   0.63331   0.63331
```

```
     2 1955.6707                    1404.5034   0.63331   0.63331
     3  315.9688                     333.8850   0.63331   0.63331
     4 1963.1671                     149.1711   0.63331   0.63331
     5  363.8099                     272.2600   0.63331   0.63331
     6 1825.9194                       0.0000   0.63331   0.63331
     7    0.0000                    1775.8244   0.63331   0.63331
     8 3163.2951                     635.6860   0.63331   0.63331
     9 1436.2990                    3832.8012   0.63331   0.63331
     :     :                            :         :         :
   223    0.0000                     121.9740   0.63331   0.63331
   224   44.8950                       0.0000   0.63331   0.63331
   225    0.0000                     156.8550   0.63331   0.63331
```

Productions and attractions are actually coded in parallel, on one line per zone, unlike the schematic representation in table E.1. Both trip end totals and their confidences are read as real numbers from their 10-column fields; the decimal point does not have to be located in a particular column. This example shows ordinary weighting with confidences set at 100/157.9.

The example shows that there are no productions from zones 7, 223 and 225, or attractions to zones 6 or 224. These are empty zones. MVESTM checks and produces a fatal error for zero trip ends and non-zero rows or columns in the prior matrix and vice versa.

While working from the full PA matrix observed by household surveys, a common set of trip ends served for both observed and synthesised datasets, since the synthetic trip distributions were constrained to the observed trip ends.

To complement actual count data without reliance on an observed trip matrix, synthesised HBW car trip ends were extracted from the WTSM model after modal split.

Different trip end files were produced with strong, ordinary and weak confidences.

E.1.4 Screenline file

A screenline file is optional, and is specified by `SCRFIL=T` in the control file. It is usual in conventional matrix estimation as it introduces count information and hence is shown on the left of table E.1.

It also defines the link on which each count is observed. In conventional applications, MVESTM processes these to determine the OD movements intercepted by each screenline. Links forming a screenline can be selected in Trips and Cube graphical interfaces to generate a screenline file, taking counts and confidences from the links' volume fields or attributes held on the network.

The screenline file is a text file, with columns for the screenline number, A and B nodes of the link crossing the screenline, the counted volume of trips, its confidence, and direction. The example below is the complete file for the observed household PA data aggregated to 3x3 segments.

```
*TRIPS Screenline format WIDEND=2 from Genstat
* for use with preformed intercept file
* Anode,Bnode columns hold Prod,Attr sectors
* Observed trips
*Scrn<-Anode-><-Bnode->< Trips  ><Conf >                    D
*        1         2         3         4         5         6
*234567890123456789012345678901234567890123456789012345678901234567890
S 101        1        1 57890.0750.63331                   1
S 102        1        2  3218.3350.63331                   1
S 103        1        3  6132.4620.63331                   1
S 201        2        1 12552.3340.63331                   1
```

```
S 202        2        2 22299.3290.63331                  1
S 203        2        3  4621.8170.63331                  1
S 301        3        1 11881.5450.63331                  1
S 302        3        2  1018.3250.63331                  1
S 303        3        3 63601.5030.63331                  1
```

There are different formats for the screenline file, involving tolls, turns (C nodes, in Cube Analyst) and screenline names. Version 2 has been adopted in this study; it is best to specify this in the control file as `WIDEND=2`.

Each screenline is identified by a unique number. Screenline numbers need not be sequential. There can be multiple records for a screenline, one for each link crossing it. Links are defined as one or two way in the final column. The records do not need to be ordered or grouped by screenline number; MVESTM can accumulate counts by screenline number.

MVESTM also aggregates the confidences coded for individual links, weighting the average by the counts. If there are only counts of zero, the confidence is set to zero. In practice, zero counts are unusual - links with low flows are rarely modelled or surveyed, but there may be zero counts of vehicles in rarer classifications. When the observed matrix is presented as screenline data for this study, there are many cells with zero counts, even excluding empty zones. These are genuine observations and useful information, obtained under the same sampling scheme and with the same confidence as non-zero cells. The MVESTM program was therefore altered to retain the confidence coded in the screenline file for a zero count.

A more sophisticated approach to aggregating confidences would allow for the inefficiency of uneven sampling (section 3.10.1 and CN 11.4)

Both counts and their confidences are read as real numbers within their fields; the decimal point does not have to be located in a particular column

In this study, the screenline file is used to introduce trip information aggregated over sets of PA movements in general, not necessarily those crossing physical screenlines or passing along any particular set of links. The definitions of these sets of movements are provided separately to MVESTM in an externally generated intercept file (see below).

Data for both screenline and intercept files were prepared in a Genstat program. This read in the observed matrix, synthesised data sets by fitting it to trip distribution models, aggregated them by segments or by intercept proportions that had been derived from the WTSM EMME/2 model, and output a screenline file in Trips format and a corresponding intercept file in a simple text format.

Where multiple movements are represented by one screenline, the Genstat programme sums the volumes, and outputs just one line in the screenline file for compactness. The file could be produced with one line per PA movement, since MVESTM aggregates by screenline. This format could also provide a text input for the intercept file, if origin and destination zones were substituted in the A and B node columns. This was tried but encountered problems using calls to existing subroutines (UINTG) to read the screenline file. The screenline file would need a column of 'proportion intercepted' to hold all the information for an intercept file, and would often be very large.

Since a separate intercept file is provided, the link definitions in the screenline file are not processed by MVESTM, but production and attraction sector numbers are placed in the A- and B-node columns as annotation. Unique screenline numbers are generated from these sector numbers.

Screenlines are listed in the order of their screenline number, for convenient correspondence with the intercept file's sequential numbering.

For calibration from the observed matrix, a different screenline file was generated for each level of aggregation, or each segment. The Genstat program wrote out four columns of trips; the observed data, and synthetic data from trip distributions modelled with Exponential, Tanner and Power deterrence functions. Unwanted columns were removed by text editor to give a different file for each of the four data sources.

For calibration from real counts, a screenline file was generated for each combination of individual, aggregate or separate counts by physical screenline, period and direction. Aggregated counts were presented on one record as one screenline. These screenlines were numbered sequentially, but the records were identified by dummy node numbers based on the count source in the Anode and Bnode columns. The screenline files were generated in Genstat in parallel with the corresponding intercept files.

E.1.5 Network with part-trip (level 2) data

Not shown in table E.1.

Later in the study, an alternative to the screenline file was found as a way of entering data. It was demonstrated, but not developed or used for the main analyses. It took advantage of MVESTM's part-route facility, sometimes available at level 2 (L2) of MVESTM's implementation and licensing, and invoked by setting `PRTTRP=T` in the control file.

This handles local numberplate or station-to-station surveys by assigning them to a network, and then treating the assigned flows as link counts for matrix estimation. Unlike counts presented in the screenline file, no corresponding parameters are generated in the model structure.

At an intermediate stage the counts are held in a volume field of a network. Their confidences are set in another volume field. An artificial network was prepared with observed trip matrix cell values in the count volume fields, to present the trip matrix as data to MVESTM.

The artificial network comprised two rings of one-way links. One ring had 225 zones attached, as for the dummy route choice probability described below. The second ring comprised 31,428 links, one for each matrix cell after excluding empty zones. The two rings were linked simply for network connectivity.

The network was built in MVNET, creating but not populating volume fields with `NEWVOL=20`. The network was run through MVESTL with a dummy survey file:

```
* Part Trip Survey Data
*   Dummy to initialise network for observed trip matrix data
*     Station Nos.        Network Numbers
*     In     Out        In Link       Out Link     Trips
*        1         2         3         4         5         6
*2345678901234567890123456789012345678901234567890123456789012345678901234567890
       1       2   41429   10001   41429   10001       0
```

This marked the network as carrying part-route data, and set the volume fields for the count (VLFLOW=1) and confidence (`VLCONF=2`). These fields were then populated with the observed trip matrix cell values using the network manipulation facility of the AVCAP program in the Trips suite.

An intercept file was prepared with one record for each 'counted' link, identifying the one PA movement that was the source of the count data in the observed trip matrix. The intercept file was keyed to the link's Anode and Bnode, instead of the screenline number as described below.

MVESTM appeared to ignore zero counts, so these were seeded. MVESTM did not appear to recognise the scaling of the volume fields by `VOLSCL`, so the seeds could not be less than 1. Allowing for this perturbation of the dataset, the expected results were achieved with a quick run time.

Entering the data as AVCAP manipulations was slow, but volume field data is awkward to present to MVNET on a second line (try the csv format?), and MVNET does not appear to update it in a compiled network, eg after running MVESTL.

A limit on link numbers was encountered below the full matrix size of 50625 (42000+?). This might be overcome by using volume fields for the eight line groups available in public transport networks.

E.1.6 Intercept file

The intercept file holds the definition of each screenline (or L2 link) as the OD movements that it comprises. The intercept file appears in the middle of table E.1 because it defines the scope both of the screenline file's count information, to the left, and of the corresponding parameter in the model structure to the right. No entry in the intercept file is needed for matrix or trip end data or parameters, because their scopes are implicit – cells, rows or columns of the matrix. Conceptually, a screenline can be defined in the intercept file as any linear combination of matrix cells.

In conventional practice, an intercept file is not usually input to MVESTM. Its contents are generated internally by identifying the OD movements that are routed along each link in a screenline file (or the surveyed L2 network). The routing information is input as a route choice probability file (see below). An intercept file may be used to restart matrix estimation from an earlier run without re-processing this routing information. It is specified by `INTCPT=T` in the control file.

The TRIPS intercept file is essentially an internal working file. It has a binary sequential format with a complex blocked structure, described in the document file InterceptFile.doc provided by Citilabs

For this study, the OD movements comprising each screenline (or L2 network link) are defined externally. They are prepared as a text file, and then converted into the TRIPS binary format with a conversion program, Text2icp.exe. This is written in Fortran 77, and calls existing MVESTM subroutines for writing out the intercept file.

The text file has columns for screenline number, production zone, attraction zone and proportion intercepted. There is one line for each PA zone pairing intercepted by a screenline. Several lines can have the same screenline number, allowing the generation of intercept files for screenlines that comprise multiple zone-to-zone movements, as is the usual case. The screenline numbers increase sequentially from 1, as in the binary file, and correspond to the numerical order of screenlines in the screenline file. All intercept proportions are unity for screenlines representing segments, which are made up of complete PA movements.

An example of the text file format is given below. It is an abstract from the definition of 3x3 segmentation, corresponding with the screenline file example.

```
*Intercept data from Genstat
* for conversion to TRIPS.icp file
* by Text2icp.exe
* 3x3 segments
* Scrn  Prod  Attr    Prop
     1     1     1    1.0000
     1     1     2    1.0000
     1     1     3    1.0000
     1     :     :     :
     1    85    83    1.0000
     1    85    84    1.0000
     1    85    85    1.0000
     2     1    86    1.0000
```

```
2     1    87    1.0000
2     1    88    1.0000
:     :     :     :
9   225   223    1.0000
9   225   224    1.0000
9   225   225    1.0000
```

Screenline number 1 in the intercept file corresponds with number 101 in the screenline file. The screenline represents the segment of movements from sector 1 (zones 1–85) to sector 1. The second screenline, numbered 102 in the screenline file, represents movements from sector 1 to sector 2 (zones 86-131). The final sector, numbered 9 in the intercept file and 303 in the screenline file, represent movements from sector 3 to sector 3 (zones 132–225).

The full length of the file is 50,625 lines – 225 production zones x 225 attraction zones – excluding the headings which are removed before input to the conversion program. The text file is generated in Genstat, alongside the screenline file.

There is a different intercept file for each level of aggregation, or for each single segment, for calibration from the observed trip matrix. Where all segments are presented, there are always 50,625 lines, because each zone-to-zone movement is included or 'intercepted' in one and only one segment.

The pattern is more complex with real screenlines. The proportions of each OD movement intercepted at each screenline by period and direction were taken from the WTSM EMME/2 assignment model, and factored and transposed to relate to a 24-hour PA matrix. Where counts were aggregated, the corresponding intercepts were aggregated. One intercept file was prepared for each screenline file in accordance with the combination of counts in the screenline file.

E.1.7 Route choice probability and path files

Not shown in table E.1.

Although route choice probability information is not needed when an intercept file is input, MVESTM still expects a file, for the correct number of zones. A dummy was generated from an artificial circular network, with one link per zone, and one zone attached at each node. A simple star network was not possible because of a limitation on the number of links attached to a node.

In a small early test, intercepts were generated in the conventional manner from a network via path and route choice probability files. The network was artificial, with a direct link between every zone pair. All links had equal lengths, speeds and hence costs, so with all-or-nothing assignment each link carried just one OD movement. Any set of OD movements could be defined by specifying their direct links as comprising a screenline, except for intrazonal movements which never appear on the network. This approach is unwieldy on a large scale.

The Trips suite was used to compile networks in MVNET, generate a path file in AVROAD (`SAVPTH=T`), and convert it to a route choice probability file in MVRCP (`RCP=T`). The Trips path and route choice probability file formats are binary. Voyager offers an alternative path file format, and can form an intercept file directly from its assignment modules, but these were not investigated.

E.1.8 Parameter file

The parameter file appears on the right of table E.1, since the parameters determine the structure of the estimated matrix. The file lists all the parameters which can be adjusted to give the optimal fit. The optimal fit is to all of the data on the left of table E.1. Each parameter may affect the fit to any data item; it does not optimise the fit just at the corresponding item of data to its left in table E.1.

Following a header line, the list always starts with row and column parameters, even if there is no corresponding trip end data. These parameters allow Furness-like fitting, and also provide for factoring of movements that are not intercepted by screenlines.

The file then lists a parameter for each screenline, in order of their numbering in the screenline file. These apply to the set of movements intercepted by the screenline. They are the usual factors applied in matrix estimation to fit to counts.

When a cost matrix is supplied, as in this study, the final two parameters are coefficients of the cost deterrence function, alpha and beta.

The parameters are numbered consecutively, starting at 1, and running in a single sequence for the trip end, screenline and cost parameters up to $2 \times N_{zone} + N_{screen} + 2$. There are four data columns, for initial values, lower and upper limits, and scaling.

Parameter files are not input in conventional practice; MVESTM can deduce the structure from the screenline file and the presence of a cost matrix, and applies default values. Although the parameter files are text files, they are intended for internal use only, and their formatting is strict. They can be edited by MVESTE, but this loses extra information in an output parameter file which is needed for a warm start. Input of a parameter file is specified by `MODPAR=T` in the control file.

The example below is an abstract from the parameter file for fitting an Exponential deterrence function to data from nine screenlines, eg the observed HIS trip data aggregated to 3x3 segments of previous examples.

```
  Model parameter file      461     225     225       9       1
(line above continued)0.0000000E+00 0.0000000E+00       0     356
       1 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       2 0.1859745E+01 0.1000000E-06 0.1000000E+11 0.1859745E+01
       3 0.3706495E+00 0.1000000E-06 0.1000000E+11 0.3706495E+00
       4 0.2000514E+01 0.1000000E-06 0.1000000E+11 0.2000514E+01
       5 0.4024337E+00 0.1000000E-06 0.1000000E+11 0.4024337E+00
       6 0.2478364E+01 0.1000000E-06 0.1000000E+11 0.2478364E+01
       7 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       8 0.2844835E+01 0.1000000E-06 0.1000000E+11 0.2844835E+01
       9 0.1127352E+01 0.1000000E-06 0.1000000E+11 0.1127352E+01
       :  :             :             :             :
     448 0.8318164E+02 0.1000000E-06 0.1000000E+11 0.8318164E+02
     449 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
     450 0.9477657E+01 0.1000000E-06 0.1000000E+11 0.9477657E+01
     451 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
     452 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
     453 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
     454 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
     455 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
     456 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
     457 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
     458 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
     459 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
     460 0.0000000E+01 0.0000000E+01 0.0000000E+01 0.1000000E+01
     461 0.5000000E-01 0.1000001E-06 0.1000000E+11 0.1000000E+01
```

The first 450 parameters are for trip ends, 225 productions and 225 attractions.

The first parameter, for zone 1 productions, is fixed by setting the lower and upper limits to the same as the initial value, 1. This avoids aliasing between production and attraction trip end parameters, and sets a reference level for them. It is a default setting in Trips.

Parameters 7 and 449 are also fixed because they are for empty zones; there are no observed productions from zone 7, or attractions to zone 224. The values are fixed to one rather than zero to avoid problems taking logarithms. The prior matrix row and column to which they apply have been set to zero. If these parameters were not fixed they would tend to limits, not contributing to the optimisation, and generating warning conditions. Fixed parameters are excluded from the optimisation process, reducing the computational burden.

The other trip end parameters have initial values set to their expected values, to improve convergence. Because the prior matrix derived from costs provides only the deterrence function, the parameter values incorporate both trip end totals and balancing factors. They are taken from the fitted coefficients of the production and attraction factors of a GLM run in Genstat. They are specific to the form of deterrence function.

The same values appear in the final column. This scales the parameter to improve the efficiency of optimisation, according to the Trips help page <MVESTM><Notes on Program Use><Selection of Model Form>. Because trip end parameters comprise both trip end totals and balancing factors in a trip distribution model, they take a relatively wide range of values, compared with the range of trip end growths encountered when updating a matrix, so there may be particular benefits in scaling here. However, the benefits to convergence and model stability are less apparent than those of providing good initial values and removing empty zones from the optimisation.

The Trips help page <MVESTE><Notes on program use><Using the editing facility> "Normalising parameter values" and <Program specific data><Model Parameter Editing Commands> (at the bottom of both pages) suggest that the reciprocal of the expected parameter value should be specified for scaling. This was tried, but gave errors.

The lower and upper limits, in the middle columns, are generally Trips defaults.

Following the trip end parameters, the next nine parameters, 451 to 459, apply to the screenlines for 3x3 segments, which are used to introduce aggregate OD trip data. A simple trip distribution model is fitted by the deterrence function and trip end parameters alone, so these screenline parameters are fixed at unity. If they were not, they would act as K factors, and would be aliased with the cost deterrence function (and with the trip end parameters, unless their reference levels were fixed, like parameter 1).

The final two parameters, 460 and 461, are the cost coefficients alpha and beta for the Power and Exponential deterrence functions, or their components of the Tanner. The example is to fit an Exponential deterrence function; the final parameter is given an initial value for its coefficient that is in the right range to aid convergence, and might be set from other studies (eg table 8.2 of Bly et al 2001), but is not a fitted value for this dataset as in the case of trip end parameters.

The coefficient of the Power component alpha, parameter 460, is fixed at zero. Zero is the null value for cost parameters, rather than unity as for other parameters. Unlike other parameters, cost coefficients can take negative values; their logarithms are not taken during pre-processing for optimisation, and an edit check was relaxed in the source code. Alpha takes a negative value for a normal Power model, where travel decreases with cost, or for the concave form of the Tanner model that fits better in Wellington.

The coding of the cost parameters for a Power deterrence function is:

```
ALPHA -0.150000E+01 -0.100002E+02 0.1000000E+02 0.1000000E+01
```

```
 BETA 0.0000000E+01 0.0000000E+01 0.0000000E+01 0.1000000E+01
```

for a Tanner deterrence function:

```
ALPHA -0.500000E+00 -0.100012E+02 0.1000000E+02 0.1000000E+01
 BETA 0.3000000E-01 0.1000011E-06 0.1000000E+11 0.1000000E+01
```

and for the flat, null model:

```
ALPHA 0.0000000E+01 0.0000000E+01 0.0000000E+01 0.1000000E+01
 BETA 0.0000000E+01 0.0000000E+01 0.0000000E+01 0.1000000E+01
```

Note variations in the last digits of the lower limits in the second column. These provide unique keys for text searches on output parameter files, for abstracting fitted values of the cost coefficients.

Input parameter files were assembled in a text editor, taking an output screenline file as a template. Different files are needed for different levels of aggregation, which affect the number of screenlines, and different deterrence functions, which affect the initial values and scaling of trip end parameters as well as cost parameters.

Output files carry extra information, including pointers which show the re-ordering of parameters to exclude from optimisation those fixed in the input or by the optimiser itself. These are used for restarting MVESTM or further statistical analysis in MVESTE. They are lost when MVESTE is used to edit the parameters.

Parameter numbers take the first eight columns in this format. Earlier formats had shorter numbers and a special version of MVESTE was provided to read the longer format.

E.2 Program coding and alterations

Citilabs supplied source code for the shell of MVESTM, allowing changes to the way in which inputs and outputs are processed. The core optimiser was supplied as compiled objects to maintain commercial confidentiality. Common files and libraries called by MVESTM were also supplied.

The code was for MVESTM version 7, modification 1.8, with library version 7.49.

A version of MVESTE 7.0.8 was supplied to accept long parameter numbers in columns 1-8 of the parameter file. Most processing was run with Cube version 4.1.2 or 4.2.3.

E.2.1 Features of program code

The programme is coded in Fortran 77. It employs some special features of that version of the language, allowing for some peculiarities of IBM implementations. A good deal of coding is devoted to working within the computing limitations of the early 1990s, in particular packing data efficiently into core computer memory (RAM), and holding intermediate workings in off-line files.

The shell code allows MVESTM to be run in various environments – in batch with control files, interactively from a console or within Cube's Application and Scenario Managers. It accepts various file formats, such as matrices and paths from Trips or other programmes in the Cube suite. There are provisions for accepting different combinations of data sources, and for part-trip (level 2) and hierarchical (level 3) estimation.

The structures in which data is passed to and from the core optimiser had to remain fixed, limiting the changes possible. Changes to the complex code were kept small and simple avoid to unintended consequences. Access to the source code was valuable for the insights it gave into the program's operation.

Amended Fortran code was recompiled with a Salford personal edition compiler from Silverfrost and linked with compiled objects using a 'make' file supplied by Citilabs.

E.2.2 Sign of parameter for Power and Tanner function

All parameters are expected to be positive; in the case of trip end and screenline parameters, this prevents negative cell values appearing in the estimated matrix, provided there are none in the prior. Apart from practical considerations, Poisson likelihood functions do not take kindly to negative counts, whether they be of matrix cells, trip ends or screenline crossings.

MVESTM has two cost parameters for gravity modelling. They are alpha and beta, in effect coefficients of Power and Exponential deterrence functions individually, or the Tanner function in combination.

$$t_{ij} = c_{ij}^{alpha} \exp(-beta \times c_{ij})$$

The Exponential parameter beta is read as a positive value, but its sign is reversed internally so trip-making tends to decrease with increasing cost.

The Power parameter alpha retains its positive value internally, so it cannot provide a sensible deterrence function on its own, with beta fixed as null. In combination with beta interpreted as a negative value, the Tanner function takes a humped form, with trip probabilities decreasing as costs become small as well as large.

This form has been shown not to fit the Wellington data well; a negative value of alpha is needed to improve on the simple Exponential model. To allow this, the check on parameters for non-negativity was been bypassed for alpha in the source code. This allowed fitting of Power and all forms of Tanner function.

E.2.3 Confidences for a zero screenline count

MVESTM's original algorithm to aggregate link counts within a screenline set the screenline confidence to zero if the total screenline count was zero. The algorithm was changed to retain the confidence coded for the last link of the screenline if all link counts were zero.

E.2.4 Limits on the number of screenlines

To represent each cell of the observed trip matrix as a separate screenline required 162x194=31428 screenlines, after excluding empty zones.

This needed more than the standard four digits for numbering screenlines in the screenline file. The shell code was altered to read a digit in the first column of the card format, in place of the card type 'S'. Together with the following four columns usually used for the screenline number, this allowed five digits.

The parameter file format had already been altered to allow larger parameter numbers.

Limitations on core space were then encountered. More memory was allocated to MVESTM (with the 'memory' button on the control file editor), but the program required more memory than could possibly be made available.

The memory limitation is not directly a consequence of trying to optimise a very large number of parameters including those for screenlines. Only variable parameters are presented to the core for optimisation. Parameters that are fixed in the parameter file, as for all screenlines in trip distribution calibration, are excluded from the optimisation problem.

The limitation seems to arise from the allocation of memory in preparation for running the optimiser. Blocks of memory whose size is proportional to the square of the number of screenlines seem to be critical. They may be used for function evaluation.

Without detailed knowledge of the optimisation algorithm, it was not possible to be sure whether the limitation is intrinsic to the method. However, it is quite plausible that it only arises as an efficient method for handling a much smaller number of screenlines in core, which could be re-arranged to deal with more

screenline data. MVESTM can accept much larger prior matrices, containing more information than was presented as screenline data. However, without access to the core optimiser, the limitation could not be overcome directly.

Later presentation of all observed trip matrix cells for non-empty zones as part-route (L2) link counts demonstrated that this amount of data is well within the capacity of the underlying routines.

E.2.5 Deviance calculations

The print file lists observed and estimated trips for origins, destinations, and screenlines, with their absolute and relative differences. These are sufficient to review the results, but for further processing the rounding lost some precision and the listing over two lines was awkward to edit.

The code was altered to append the deviance to the end of each line in scientific format, to show very small differences in fit. The deviances were also totalled and printed out for origins, destinations and screenlines.

Provision was made for calculating expected deviances, but since the aggregate data was generally non-sparse, this was left as unity, and just appeared as a count of units.

Deviances for trip ends were also calculated from the input trip ends and estimated matrix in a Voyager MATRIX module, but screenlines could not be processed in the same way.

E.2.6 Hessian listing and calculation of parameter variances

The Hessian matrix was listed in the print file as a diagnostic when investigating the sensitivities and standard errors output by MVESTE for a small test network.

For the full WTSM matrix, this produced very large print files, over 5MB, and so was suppressed. An attempt was made to calculate the standard errors and covariance of the cost coefficients alpha and beta from the Hessian. However, the re-ordering of the Hessian to exclude parameters fixed internally and externally could not be established with confidence.

An attempt was also made to estimate the standard error of cost parameters by perturbing their fitted values, and re-running MVESTM with the perturbed values fixed. The parameter file was edited programmatically with MVESTE.

The format for summarising confidences in the print file was extended to show the full precision of the smaller real confidences based on sampling rates etc.

E.3 Managing multiple scenarios

The work described above enabled the calibration of trip distributions by MVESTM on a full-scale model like the WTSM. It also presented many options for testing, both in the basic calibration from the observed HIS matrix:

- segment size – aggregation to 3, 6, 10, 15, or 65 sectors
- choice of individual segment
- source data – observed or fitted with different deterrence functions
- weighting of trip ends – strong, ordinary, or weak
- deterrence function – Exponential, Tanner, Power, or flat as a null model

and in model estimation from traffic counts, aggregated or separated by:

- screenline
- period
- direction.

With so many options, it is convenient to run them in batches, rather than one at a time. MVESTM and supporting programs can be quite profligate in their use of files, so it is important to identify and preserve key input and output files for each option systematically.

Development was initially undertaken in Citilab's Application Manager, a graphical environment. It was supplied with Trips, and was previously known as TRIPSWIN. It has some facilities (application versions) for handling different scenarios, but these have been much improved and largely superseded by the Scenario Manager in Cube Base. Citilabs kindly granted a license for Cube Base. This is based on catalogs of data and applications, with keys that can be varied between scenarios.

E.3.1 Inputs

Keys often define whole files, but they can also define numbers or strings that are substituted in files. At one stage keys were inserted in screenline and parameter files for input to MVESTM. These files do not incorporate keys as standard, and require exact substitution of strings, particularly in the parameter file. Column 63 of the `&FILES` block in the resource file also has to be edited – see <Cube Base><Application Manager><Application Manager resource files><Resource file formats><Format specification for Cube Voyager, TP+, TRIPS program resource files> in Cube Help.

Because the size and content of these files varies considerably between scenarios, later practice was to prepare all files individually beforehand, and include the appropriate catalog keys in the file names. Input files are then linked to catalog keys, so the appropriate file is run in each scenario. File naming keys were substituted in Trips MVMODL control files for input matrix preparation, and Voyager MATRIX script files for processing the outputs.

E.3.2 Outputs

Important output files were marked as scenario specific, so they could be retained in a directory structure defined by the scenario. The scenario was also included in the file name. Scenario naming followed the values of the catalog keys that defined the scenario.

Catalogs were generated and run for just one level of trip end confidence. After running at one of the levels, the 'Base' directory (with all its sub-directories) was renamed 'Stng', 'Ordy' or 'Weak' according to the level of confidence used, and the catalog was run with another level of confidence into a new Base directory. This was a convenient method for managing old or trial catalog runs.

It was found helpful to have the same length for all the alternative strings for a key that formed part of a directory or file name. For example, the keys for deterrence models were 'Exp', 'Tan', 'Pow', or 'Flt' (for a flat, null model without cost deterrence). This allowed easier sorting of text abstracts from the files, or searching with wild cards such as '?'. Alphabetical ordering of the strings in a logical order (eg sector size) was also helpful.

The usual set of scenario specific MVESTM output files was:

Print file	*MV*EST*00A**.prn
Estimated matrix	*MV*EST*00A**.mat

Model parameter file	*MV*ESM*00A**.dat
Execution log	*MV*ESL*00A**.dat

The estimated matrix was compared with other matrices in a Voyage MATRIX module, from which the following output files were kept:

Print file	*MV*MAT*00A**.prn
Record file of the estimated matrix	MAT*.dbf
Print file of trip ends cumulated in	CuTripEndFit.csv (in Base directory)

The MATRIX module also output a difference matrix file in Trips format to display both positive and negative desire lines and trip ends in MVGRAF. Separate positive and negative matrices have to be presented to do this in Cube's network window. MVPRIN and the Cube matrix window were also used to examine matrices.

In a large catalog with many scenarios, results were spread over many directories and files. All the scenario-specific values for an output of interest were consolidated into a single text file by <Search> <Find in Files…> of the UltraEdit text editor. The values were found by searching on key strings such as special upper and lower limits for cost coefficients in parameter files. These summary text files were edited and sorted for import into Excel spreadsheets, or sometimes Access relational databases. For statistical processing these were read into Genstat.

E.4 Run times

On a Toshiba laptop with an Intel T7200 2GHz processor and 2GB RAM running under Windows XP SP3, estimation of a 225 zone internal WTSM matrix usually took less than five minutes, and often less than one minute.

The part-trip L2 formulation, with individual cell values of the observed matrix as data, but without parameters corresponding to them, seemed to run quicker.

MVESTM has special hierarchical methods (level 3) for handling large matrices. They might have particular application for processing a hypermatrix representation of matrices with multiple dimensions, but no formulation was found for applying this method to fitting multivariate generation models or period and direction factors.

Appendix F: MVESTE statistical problems

F.1 Small-scale matrix estimation

While developing the use of MVESTM within Cube, the opportunity was taken to run very small but well controlled models to investigate the measures of fit and accuracy output by the MVESTM suite, in particular the statisitical outputs provided by the MVESTE program..

The models comprised two zones with two one-way routes between them, allowing one unique route for each direction of movement. The models generally omitted cost information and hence gravity modelling, and provided only one or two pieces of information about a movement, from

- prior matrix cells
- zonal trip ends, or
- screenline counts.

Only one parameter was allowed to vary to fit the data for each movement, to produce a predictable result. Other parameters were fixed as null and other information sources were ignored by setting their confidences to zero.

For such special cases, there are expected to be simple relationships between the count information input, the Hessian matrix of the parameter estimates, the standard error of the parameters produced by MVESTE, and the sensitivity matrix produced by MVESTE in an alternative mode, in particular:

$$1/(\text{Standard Error})^2 = \text{Diagonal(Hessian)} = \text{Sensitivity} \times t^2$$

where t are the trips in the prior matrix.

The Hessian is the matrix of second differentials of the objective or likelihood function with respect to the parameters. It is important in the search to optimise the objective function or maximise the likelihood, and in estimating accuracies. MVESTM usually outputs it only in the internal gradient search (.GDS) file in binary format, but the source code has been altered to output it as text, giving access to this important intermediate working stage.

The optimisation is an iterative procedure, and calculation of the Hessian is computationally demanding. Default settings for MVESTM limit the recalculation of the Hessian to reduce computation time, though this is less critical with modern computers. Some problems with measures of accuracy were resolved by forcing recalculation of the Hessian, either by setting `ITERH=1` in MVESTM, or by setting `RECALC=T` in MVESTE for calculating parameter standard errors. This does not affect MVESTE when calculating matrix sensitivities; in this mode, the Hessian from MVESTM in the gradient search file is not used by MVESTE either.

Three features in the measures of accuracy remain unresolved:

1 Parameter standard errors are close to expectations for large amounts of count information, but are larger than expected for small counts, drastically so for counts much less than 1. This might affect trip distribution models which are calibrated from sparse matrices.

2 Elements of the Hessian returned by the core optimiser do not respond to changes in the scale of the parameter as would be expected. This can occur with changes in a prior matrix with no information value (confidence=0); its cell values still scale the output matrix. Values of the Hessian are as expected when parameters are equal to 1, ie the prior matrix is the same as the estimated.

3 The sensitivity matrix gives unexpected results for a reverse movement, which is identical to but independent of a forward movement, for which the results are as expected. As with the Hessian, this difference arises when parameters are not equal to 1.

These findings apply to a very limited set of tests.

The value of the objective function, FBEST in the execution log file, has been checked against the Poisson likelihood by manual calculation, and found to correspond. It differs from the deviance of GLMs only by a scaling factor that is a function of the data. Expectations for the Hessian have been checked by finite difference methods, calculated from changes in the objective function as (fixed) parameters are perturbed about the best fit.

All tests produced the expected values for parameters and hence the estimated trip matrices. This includes variants with:

- multiple and conflicting information sources
- varied confidences
- information and/or parameters taken from
  - prior matrix cells
  - zonal trip ends, or
  - screenlines.

Together with the objective function corresponding with the prime measure of fit used in the rest of this study, this gives confidence in the functioning of the core optimiser, even if measures of fit produced by auxiliary programs are not understood.

F.2 Full-scale model calibration

Later attempts were made to calculate accuracies of parameters, particularly cost coefficients, after calibration of trip distributions from full scale Wellington data.

The confidence ranges given by MVESTE appear large, often greater than the parameter value, and overflowing their print formats. This is at odds with the high significance of cost effects in trip distribution. Accuracies appear to be affected by parameter scalings (in the 4th column of the model parameter file), which should be neutral.

MVESTE sometimes produced untrapped system errors. These may occur when the optimiser has fixed parameters. MVESTE's reporting of such fixed parameters could differ from that of the optimiser in MVESTM. The performance of MVESTE could differ markedly if the Hessian was recalculated (RECALC=T).

The fixing of parameters by the optimiser appears to be affected by the size of confidences (including cost) even when the relative sizes are maintained. The Hessian returned from the optimiser appears to be in the same order (ASSOC) as parameters are input by the MVESTM shell, not ordered by IPOINT as suggested in notes in the source code of MEST6.FOR. The correlations with fixed parameters in the Hessian are not all zero.

Appendix G: Spiess's Poisson errors with trip end constraint

Spiess, H (1987) A maximum likelihood model for estimating origin-destination matrices. *Transportation Research B21,* no.5: 395–412. Section 4: The matrix estimation problem with given marginal totals.

Spiess considers maximum likelihood solutions for Poisson errors in an observed trip matrix, giving a 4×4 matrix with fixed trip end constraints as a worked example. This can be seen as a special case of matrix estimation with trip ends rather than links counted, and imposed as fixed constraints. Each movement is counted exactly twice. Under maximum entropy, the solution would be of the multiplicative form:

$T_{ij} = \alpha_i \beta_j t_{ij}$ where t is the observed matrix and T is the estimated matrix

but with a Poisson distribution of t, maximum likelihood is shown to result in the form:

$T_{ij} = t_{ij} / (\rho_{ij} + \alpha_i + \beta_j)$ for $t_{ij} > 0$, where ρ_{ij} is the sampling fraction

and $T_{ij} > 0$ for some $t_{ij} = 0$

G.1 Example

The example matrix and trip end constraints are shown in table G.1. Neither the individual trip end constraints nor their total match those of the observed matrix. The sampling fraction is the same for all observations.

Table G.1 Observed matrix and trip end constraints

i\j	1	2	3	4	Total	Constraint
1	10	0	15	5	30	24
2	0	11	0	0	11	18
3	6	9	5	0	20	19
4	0	0	8	4	12	11
Total	16	20	28	9	73	
Constraint	19	17	23	13		72

Source: Spiess (1987), table 1

G.2 Fitting by MVESTM

The example was solved using MVESTM with the following formulation and coding. File coding is discussed in more detail in appendix E, section E.1.

Table G.2 MVESTM formulation

Data	Scope	Structure
Prior matrix		
Unit matrix, zero confidence		$t_{ij} = 1$
Trip ends		**Parameters**
Constraining trip ends 24 : 11 19 : 13 Confidences high (1,000,000) to represent constraint	origins destinations	*Fixed*
Screenlines	**Intercepts**	
Observed OD trip data 10 : : 4 Confidence = 100	One ij cell per screenline, all i,j	Free to fit posterior matrix cell values
		No cost parameters

Italic – null effect on fitted model

The trip end file holds the trip end constraints with very large confidences to prevent any changes in them. The corresponding trip end parameters are fixed at unity.

A separate screenline is set up for each cell, both to present the trips observed for the cell as data and to generate a separate parameter to estimate the cell value in the output matrix.

The input matrix is a unit matrix, to provide a base for the screenline parameters. The confidences are all zero so it is not regarded as data.

The observed trip data could have been introduced through the input matrix file, but would still have needed screenline parameters (with null data, confidence set to zero) to fit each cell individually. The parameters would then be multiplying the observed OD trip data, rather than the unit matrix, and could not estimate a non-zero output cell where there is a zero observation.

G.2.1 Control file

```
Matrix estimation
&FILES
IMAT1='SpiessUnit.MAT',
IDAT1='SpiessTripEnd4.DAT',
IDAT2='Spiess16.prm',
IDAT4='Spiess16.scn',
IDAT5='Spiess16.icp',
IDAF1='Dummy4zone.RCP',
```

```
OPRN= 'SPEST00A.PRN',
OMAT1='SPEST00A.MAT',
ODAT1='SPESM00A.DAT',
ODAT2='SPESG00A.GDS',
ODAT3='SPESL00A.DAT',
 &END
 &PARAM
WIDEND=2,
DEC='D',
MFORM=2,
 &END
 &OPTION
TRIPM=T,
COSTM=F,
SCRFIL=T,
TRPEND=T,
INTCPT=T,
MODPAR=T,
WARMST=F,
 &END
```

G.2.2 Input matrix file

```
IMAT1='SpiessUnit.MAT'
```

The trip matrix (101) has 1 in every cell, the trip confidence matrix (102) has 0.

G.2.3 Trip end file

```
IDAT1='SpiessTripEnd4.DAT'
```

The trip end constraints are coded with very high confidences.

```
* 4x4 Poisson with trip end constraints
* Spiess A MAXIMUM LIKELIHOOD MODEL FOR ESTIMATING OD MATRICES
* Transpn Res-B Vol. 21B, No. 5, pp. 395-412, 1987
* Section 4. THE MATRIX ESTIMATION PROBLEM WITH GIVEN MARGINAL TOTALS
* Table 1: A small example
         1   24                           19      1000000   1000000
         2   18                           17      1000000   1000000
         3   19                           23      1000000   1000000
         4   11                           13      1000000   1000000
```

G.2.4 Screenline file

```
IDAT4='Spiess16.scn'
```

One screenline is entered for each matrix cell, with the observed trips and the default confidence.

```
*TRIPS Screenline format WIDEND=2
* for use with preformed intercept file
* Anode,Bnode columns hold Prod,Attr sectors
* 4x4 Poisson with trip end constraints
* Spiess A MAXIMUM LIKELIHOOD MODEL FOR ESTIMATING OD MATRICES
* Transpn Res-B Vol. 21B, No. 5, pp. 395-412, 1987
* Section 4. THE MATRIX ESTIMATION PROBLEM WITH GIVEN MARGINAL TOTALS
* Table 1: A small example
*
*Scrn        i        j     Trips   Conf                  D
```

```
*        1         2         3         4         5         6
*2345678901234567890123456789012345678901234567890123456789012
S 101        1        1        10    100                    1
S 102        1        2         0    100                    1
S 103        1        3        15    100                    1
S 104        1        4         5    100                    1
S 201        2        1         0    100                    1
S 202        2        2        11    100                    1
S 203        2        3         0    100                    1
S 204        2        4         0    100                    1
S 301        3        1         6    100                    1
S 302        3        2         9    100                    1
S 303        3        3         5    100                    1
S 304        3        4         0    100                    1
S 401        4        1         0    100                    1
S 402        4        2         0    100                    1
S 403        4        3         8    100                    1
S 404        4        4         4    100                    1
```

The i and j columns are for annotation only, and are not read by MVESTM. The relationship between screenlines and ij movements is defined in the intercept file.

G.2.5 Intercept file

```
IDAT5='Spiess16.icp'
```

This shows the single ij movement which each screenline represents. This text file is converted to a binary (*.icp) format for input to MVESTM.

```
*Intercept data from Genstat
* for conversion to TRIPS.icp file
* by Text2icp.exe
* 4x4 matrix cells for Spiess
* Scrn     i     j      Prop
     1     1     1    1.0000
     2     1     2    1.0000
     3     1     3    1.0000
     4     1     4    1.0000
     5     2     1    1.0000
     6     2     2    1.0000
     7     2     3    1.0000
     8     2     4    1.0000
     9     3     1    1.0000
    10     3     2    1.0000
    11     3     3    1.0000
    12     3     4    1.0000
    13     4     1    1.0000
    14     4     2    1.0000
    15     4     3    1.0000
    16     4     4    1.0000
```

G.2.6 Input parameter file

```
IDAT2='Spiess16.prm'
```

The first eight parameters for trip ends are fixed by setting both the lower and upper limits in the second and third data columns to unity.

The remaining 16 screenline parameters are left free to fit a separate value for each matrix cell.

```
  Model parameter file       24       4       4      16       0
(line above continued)0.0000000E+00 0.0000000E+00       0      16
       1 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       2 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       3 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       4 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       5 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       6 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       7 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       8 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       9 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      10 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      11 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      12 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      13 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      14 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      15 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      16 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      17 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      18 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      19 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      20 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      21 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      22 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      23 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      24 0.1000000E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
```

G.3 Outputs

Table G.3 Spiess' optimal solution

i\j	1	2	3	4	Total
1	8.718	.000	10.924	4.359	24.000
2	1.958	11.606	.000	4.436	18.000
3	8.326	5.394	5.281	.000	19.000
4	.000	.000	6.795	4.205	11.000
Total	19.000	17.000	23.000	13.000	72.000

Source: Spiess (1987, table 3)

Note the presence of trips between i=2 and j=1 or 4 where none existed in the prior.

Table G.4 MVESTM output matrix

OMAT1='SPEST00A.MAT'

i\j	1	2	3	4	Total
1	8.702	0	10.949	4.351	24.001
2	1.990	11.611	0	4.398	17.999
3	8.308	5.390	5.302	0	19.000
4	0	0	6.749	4.251	11.000
Total	19.000	17.001	23.000	13.000	72.000

Zeros are actually very small numbers, c.10^{-6}, mostly on the limits for the parameters.

The correspondence with Spiess's solution is recognisable, and good enough for practical purposes, but not particularly precise. Other combinations of confidences generally gave worse accuracies, but no other settings for the MVESTM optimiser were tried.

G.3.1 Output parameter file

```
ODAT1='SPESM00A.DAT'(extra columns omitted)
```

The elements of the estimated matrix appear in this file as the fitted values, in the first data column, of the last 16 parameters, for the screenlines.

```
  Model parameter file       24       4       4      16       0
(line above continued)-0.1602488E+08 0.3611402E-04      75       9
       1 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       2 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       3 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       4 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       5 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       6 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       7 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       8 0.1000000E+01 0.1000000E+01 0.1000000E+01 0.1000000E+01
       9 0.8701701E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      10 0.1000000E-06 0.1000000E-06 0.1000000E+11 0.1000000E+01
      11 0.1094859E+02 0.1000000E-06 0.1000000E+11 0.1000000E+01
      12 0.4350741E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      13 0.1990092E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      14 0.1161075E+02 0.1000000E-06 0.1000000E+11 0.1000000E+01
      15 0.1000000E-06 0.1000000E-06 0.1000000E+11 0.1000000E+01
      16 0.4397861E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      17 0.8307675E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      18 0.5390385E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      19 0.5301944E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      20 0.1000000E-06 0.1000000E-06 0.1000000E+11 0.1000000E+01
      21 0.6905130E-06 0.1000000E-06 0.1000000E+11 0.1000000E+01
      22 0.1000000E-06 0.1000000E-06 0.1000000E+11 0.1000000E+01
      23 0.6749331E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
      24 0.4251033E+01 0.1000000E-06 0.1000000E+11 0.1000000E+01
```

G.3.2 Print file

```
OPRN= 'SPEST00A.PRN' (abstracts)
```

This shows that the trip end constraints have been maintained. The elements of the output matrix appear as the estimated screenlines counts.

```
PROGRAM - ANLYST
SUMMARY FOR OUTPUT MATRIX <Estimated Matrix    >

 MATRIX TOTAL                =       72.000103
 INTRAZONALS                 =       29.865430
 NUMBER OF NON-ZERO CELLS    =              16
 MINIMUM NON-ZERO CELL VALUE =        0.000000
 MAXIMUM NON-ZERO CELL VALUE =       11.610751

  REPORTING PRIOR/ESTIMATED MATRIX TOTALS
```

CONFIDENCE	PRIOR	ESTIMATED	ESTM-PRIOR	(ESTM-PRIOR)/PRIOR(%)
0.0	16.0	72.0	56.0	350.0%

REPORTING OBSERVED/ESTIMATED GENERATIONS AND ATTRACTIONS

GENERATIONS

ZONE NO	CONFIDENCE	OBSERVED	ESTIMATED	ESTM-OBSV	(ESTM-OBSV)/OBSV(%)
1	1000000.0	24.0	24.0	0.0	0.0%
2	1000000.0	18.0	18.0	0.0	0.0%
3	1000000.0	19.0	19.0	0.0	0.0%
4	1000000.0	11.0	11.0	0.0	0.0%

ATTRACTIONS

ZONE NO	CONFIDENCE	OBSERVED	ESTIMATED	ESTM-OBSV	(ESTM-OBSV)/OBSV(%)
1	1000000.0	19.0	19.0	0.0	0.0%
2	1000000.0	17.0	17.0	0.0	0.0%
3	1000000.0	23.0	23.0	0.0	0.0%
4	1000000.0	13.0	13.0	0.0	0.0%

REPORTING OBSERVED/ESTIMATED SCREEN LINE COUNTS

SCREENLINE	CONFIDENCE	OBSERVED	ESTIMATED	ESTM-OBSV	NO OF ODs
101	100.0	10.0	8.7	-1.3	1
102	100.0	0.0	0.0	0.0	1
103	100.0	15.0	10.9	-4.1	1
104	100.0	5.0	4.4	-0.6	1
201	100.0	0.0	2.0	2.0	1
202	100.0	11.0	11.6	0.6	1
203	100.0	0.0	0.0	0.0	1
204	100.0	0.0	4.4	4.4	1
301	100.0	6.0	8.3	2.3	1
302	100.0	9.0	5.4	-3.6	1
303	100.0	5.0	5.3	0.3	1
304	100.0	0.0	0.0	0.0	1
401	100.0	0.0	0.0	0.0	1
402	100.0	0.0	0.0	0.0	1
403	100.0	8.0	6.7	-1.3	1
404	100.0	4.0	4.3	0.3	1

PROGRAM ANLYST FINISHED WITH RETURN CODE OF 4 AT TIME 20:18:13

Appendix H: Equivalence of maximum Poisson likelihood and maximum entropy under Bell's approximation to GLS

Bell (1984) shows that if $f(T)$ is an objective function for a posterior matrix T to be similar to a prior matrix t, then by Taylor series expansion of $f(T)$ about t, a similar objective is provided by the squared errors $(T-t)^2$ weighted by $d^2f(T)/dT^2$, which is soluble by generalised least squares (GLS).

For **maximum entropy**, maximise:

$f(T)$	$= -\Sigma[\ T\ln(T/t) - T + t\ ]$	from equation 12.21, (Ortuzar and Willumsen 1994)
	$= 0$	at $T = t$
$df(T)/dT$	$= -\Sigma[\ \ln(T/t) + 1 - 1 + 0\ ]$	
	$= -\Sigma[\ \ln(T/t)\ ]$	
	$= 0$	at $T = t$
$d^2f(T)/dT^2$	$= -\Sigma[\ 1/T\ ]$	
	$= -\Sigma[\ 1/t\ ]$	at $T = t$

For **maximum Poisson likelihood**, maximise:

$p(T)$	$= \Pi\ [\ T^t e^{-T} / t!\ ]$	
	or	
$f(T)$	$= \Sigma[\ t\ln(T) - T - \ln(t!)\ ]$	taking logs
$df(T)/dT$	$= \Sigma[\ t/T - 1 + 0\ ]$	
	$= \Sigma[\ t/T - 1\ ]$	
	$= 0$	at $T = t$
$d^2f(T)/dT^2$	$= \Sigma[\ t/-T^2 - 0\ ]$	
	$= -\Sigma[\ t/T^2\ ]$	
	$= -\Sigma[\ 1/t\ ]$	at $T=t$

Thus for T = t, maximum entropy and maximum Poisson likelihood are equivalent under this approximation to a GLS model.

Appendix I: Interpretation of GEH validation criteria as MVESTM confidences

These are first order approximations for estimating MVESTM screenline confidences from assignment validation criteria. They make simplifying assumptions sufficient to derive a simple relationship.

M is the modelled value

O is the observed value

GEH is $G = \sqrt{((M-O)^2 / ((M+O)/2))}$

We define a mean $\mu = E(M) = E(O)$

$= E((M+O)/2)$ the denominator of G^2

μ, like M and O, is expressed as hourly flows for EEM and DMRB criteria.

There is a difference between M and O, but no expectation of bias,

ie $E(M-O) = 0$

We define a variance $\sigma^2 = E((M-O)^2)$ expectation of the numerator of G^2

If O is observed as a count of random arrivals, sampled with a fraction w_O such that the expectation of the sampled count is $w_o\mu$, then the sampled count will be distributed as Poisson:

$$w_oO \sim \text{Poisson}(w_o\mu)$$

If $w_o\mu$ is large (>>1), then the normal is a good approximation to the Poisson,

hence $w_oO \sim N(w_o\mu, w_o\mu)$

or dividing by w_O, $O \sim N(\mu, \mu/w_o)$

This approximate distribution of O is still Poisson-like in that the variance of the distribution, μ/w_O, is still proportional to its mean μ. It is overdispersed by $1/w_O$, which can also be seen as the expansion factor from the survey count to the hourly flow.

An overdispersed Poisson need not arise from expanding a count of random incidents; it is unlikely that this is the prime source of error in O, let alone the modelling of M. We specify this form of relationship between mean and variance, in particular

$$M \sim N(\mu, \mu/w_M)$$

Taking the difference, $M-O \sim N(\mu-\mu, \mu/w)$

$\sim N(0, \mu/w)$ since there is no expectation of bias

and $1/w = 1/w_O + 1/w_M$ if the errors in O and M are independent, but the form of the distribution remains a reasonable approximation under correlations between them.

Equating expectations of the square of the difference (M-O)

$$\sigma^2 = \mu/w$$

$(M-O)\sqrt{(w/\mu)} \sim N(0,1)$ multiplying up by $\sqrt{w/\mu}$

$(M-O)\sqrt{(w/\mu)} = (M-O)\sqrt{(w/((M+O)/2))}$ substituting $\mu \approx (M+O)/2$

$$= \sqrt{w}\, G_n$$

where G_n is the natural GEH, taking the sign of (M–O)

Hence $\sqrt{w}\, G_n \sim N(0,1)$

Alternatively $wG^2 \sim \chi^2_1$

just from the relationship between normal and chi-squared distributions. However, the two tails of the normal are equivalent to the upper tail of the chi-squared.

I.1 Derivation of w and MVESTM confidence from validation criteria

From the relationship with the normal distribution found above, we find w such that the critical value of GEH, G_{crit}, is achieved with the probability P set out in EEM or DMRB. Equating normal standard deviates for the critical case

$$\sqrt{w}G_{crit} = \Phi^{-1}((1+P)/2)$$

$$w = (\Phi^{-1}((1+P)/2) \,/\, G_{crit})^2$$

$$\text{MVESTM confidence} = 100w = 100(\Phi^{-1}((1+P)/2) \,/\, G_{crit})^2$$

$\Phi^{-1}(P)$ is the inverse of the lower cumulative distribution of the standardised normal distribution, returning the standard deviate up to which the integral of the normal is P.

$\Phi^{-1}((1+P)/2)$ returns the standard deviate which leaves two tails outside the probability P.

I.2 Implication of Poisson-like distribution by specifying criteria in GEH

This distribution of G depends on w alone, and is not a function of the flow μ. This depends on the Poisson-like variance function:

$$\sigma^2 = \mu/w$$

otherwise the GEH is not non-dimensional in μ.

If the variance function takes the

normal $\sigma^2 =$ constant

gamma $\sigma^2 \propto \mu^2$

or other general form $\sigma^2 = \phi V(\mu)$ in GLM terminology (eg Lee et al 2006)

then the distribution of G depends on the flow μ. However, this is not specified in the criteria for the GEH in EEM or DMRB, suggesting a Poisson-like variance function in the difference between modelled and observed values. (DMRB does specify flow ranges for other criteria based on percentages.)

Alternatively, any other variance function would imply that the probability of rejecting a difference arising from the accepted errors in modelling would vary with the flow. This is conceivable, but obtuse; it posits a sophistication not apparent in EEM or DMRB's validation criteria. Amongst such blunt instruments, Occam's razor wins.

More practically, an assumption of a Poisson-like variance function allows a simple interpretation of validation criteria in terms compatible with MVESTM, leading to convenience in analysis and clarity in interpretation. This is as good a reason as any for choosing a Poisson-like variance function to interpret criteria based on the GEH.

I.3 Largeness of numbers for approximation

The approximation of the Poisson to the normal depends on $w\mu >> 1$. This can be expressed as $\mu >> 1/w$, or the mean hourly flow being large compared with the overdispersion.

Minimum observed hourly directional flow by period at any of the four WTSM screenlines was 37vph at the Rimutaka screenline, interpeak (table 8.30). The average across all screenlines, directions and periods was 1472vph.

Critieria 1 and 2 in DMRB, table 4.2 suggest that flows of less than 700vph are low.

If sampling of a Poisson process, and subsequent factoring to an hourly rate, is the source of error, the overdispersion will depend on the duration of count.

Overdispersion $1/w = 1$ for a 1 hour count (minimum likely)

or $= 1/7$ for a 7 hour IP count (more likely)

Interpretation of GEH validation criteria in table 8.5 gives overdispersions between 2.4 and 35. Analysis of WTSM screenline counts by HGLM gives an overdispersion of around 16.5, which was adopted for the analysis. This HGLM analysis also supports the Poisson-like variance function, $\sigma^2 \propto \mu$.

These overdispersions are small compared with most flows, particularly those overdispersions based on sampled counting, which provide a rationale for a true Poisson distribution.

The large number approximation only matters if, say, raw observations are truly Poisson-distributed.

$$w_o O \sim \text{Poisson}(w_o \mu)$$

It is at least as plausible that the differences between model and observation are Normally distributed,

$$M\text{-}O \sim N(0, \mu/w)$$

with a Poisson-like variance function. (Model estimates will effectively be continuously distributed.) This would not depend on any large-number approximation.

I.4 Interpretations of w

- Weight
- Sampling fraction
- MVESTM confidence / 100
- 1/Expansion or grossing factor
- 1/Overdispersion
- $1/\phi$, where ϕ is the dispersion of the GLM variance $V(\mu)$

See MVESTM Help 'Mathematical summary', equation 6; similar, but never quite the same, explanation of λ.

Appendix J: Glossary

Alogit	Software for fitting choice models
AM	Morning peak period; 7am to 9am in WTSM (TN19.1)
ANOVA	Analysis of variance
AVCAP	The capacity restraint program in the Trips suite, allowing network manipulation
Biogeme	Software for fitting choice models
CBD	Central business district
CN	Course note(s) on matrix building, estimation and validation, delivered by Prof H Kirby and others at Leeds University and dated variously up to 1992. Kirby et al (1992)
Confidence	In MVESTM, a measure of data accuracy equivalent to 100xweight in a Poisson log-linear GLM
CPU	Central processor unit
csv	Comma separated variables, data file format
CSTM	Central Scotland Transport Model
Cube	A suite of transport modelling software from Citilabs; www.citilabs.com/products/cube
deviance	$-2\times$log-likelihood; a statistical measure of fit for GLMs, equivalent to the sum of squared residuals in simple regression or analysis of variance.
df	Degrees of freedom
DMRB	*Design manual for roads and bridges, volume 12* (UK) (Highways Agency 1996)
DMS	(Trip) distribution and mode split
E	Exponent in scientific number notation; expectation in probability
EB	Employer's business, a trip purpose
EEM	*Economic evaluation manual* (NZTA 2010)
EMME/2	Software platform on which WTSM runs: www.inro.ca
ERICA	Software for building origin-destination matrices from trip records, merging them with variance weighting. Developed by the Peter Davison Consultancy for the UK Department for Transport
EVA	Erzeugung, Verteilung & Aufteilung (production, distribution and mode choice), a VISUM model
Fortran	Formula translation language, a computer programming language
Fratar	Iterative process for adjusting a matrix to fit its trip ends, or marginal totals (US usage)
Furness	Iterative process for adjusting a matrix to fit its trip ends, or marginal totals (UK usage)
GB	Gigabyte
GEE	General estimating equation
GEH	A measure comparing observed and modelled flows, suggested by GE Havers
Genstat	Software package used for statistical modelling: www.vsni.co.uk/products/genstat/, including 30-day trial download.
GEV	Generalised extreme value
GHz	GigaHertz
GIS	Geographic information system
GLIM	Software package for fitting GLMs; a subset of Genstat
GLM	Generalised linear model, a specialised form of regression
GLMM	Generalised linear mixed model
GW	Greater Wellington; brand name of Wellington Regional Council, owners of the WTSM

GWR	Geographically weighted regression
HBEd	Home-based education, a trip purpose
HBSh	Home-based shopping, a trip purpose
HBW	Home-based work, a trip purpose
HCV	Heavy commercial vehicle
HGLM	Hierarchical generalised linear model
HIS	Household interview survey
IBM	International business machines
IP	Interpeak (cf AM and PM)
K&H	Kapiti and Hutt, one sector in a division of the study area into two sectors, the other being City
KALIBRI	Function for calibrating trip distributions in the VISUM package
km/h	Kilometres per hour
L2	Level 2 in MVESTM, part-route OD data from numberplating, ticketing, toll road on-to-off etc
LP	Linear programming
LU	Land use
Matrix	In transport modelling, a two-dimensional array holding a value for every combination of origin zone and destination zone. The values are usually the volumes of movement between the zones (trips per unit time), but may be other values such as costs of travel between the zones. Sometimes called trip or interchange tables
MAUP	Modifiable area unit problem, (unwanted) effects of sizing or aggregating zones
mf	EMME/2 matrix nomenclature
ML	Maximum likelihood
MVA	Martin Vorhees Associates, or Martin & Vorhees (see also MVESTM and MVGRAM)
MVESTE	Statistical add-on to MVESTM (see appendix F)
MVESTL	Part-route L2 add-on to MVESTM
MVESTM	Matrix estimation computer programme, originally from MVA now Analyst in the Cube suite from Citilabs
MVGRAM	Gravity modelling programme, ex MVA
MYSTIC	The Methodology and Evaluation Framework for Modelling Passenger and Freight on Transport Infrastructure Scenarios using the European Transport Policy Information System
$N(\mu,\sigma^2)$	Normal distribution with mean μ and variance σ^2
NHB	Non-home-based, a trip purpose
NHBO	Non-home-based other, a trip purpose
NZTA	New Zealand Transport Agency
O(n)	Order n, of splines and polynomial curves
OD	Origin (start) – destination (finish) (see also 'Trip end')
OECD	Organisation for Economic Cooperation and Development
OmniTrans	Transport modelling software: omnitrans-international.com/en
OS	Operating system
PA	Production (home) – attraction (work) (see also 'Trip end')
PC	Personal computer
Pija	Proportion of trips from i to j using link a. SATURN terminology for the intercept proportion
PM	Evening peak period: 4pm to 6pm in WTSM (TN19.1)
PRISM	Policy Responsive Integrated Strategy Model; transport model of the West Midlands, UK

PT	Public transport – bus, train and ferry
PTV	Suppliers of VISUM
R+	Statistical package
RAM	Random access memory
RDMVAR	Matrix building program of the ROADWAY suite with an option to construct an index of dispersion matrix. CN9, DMRB vol 12, section 1, part 1 (TAM) paras 2.2.19 and 8.2.7
REML	Residual or restricted maximum likelihood
RHTM	Regional Highways Traffic Model (UK), which used partial matrix infilling methods with RSIs.
RMS	Root mean square
ROADWAY	Suite of programs developed in 1980 for the then UK Department of Transport, no longer supported.
RSI	Roadside interview survey
RUM	Random utility maximisation
SATURN	Simulation and Assignment of Traffic to Urban Networks: a junction-based traffic model developed at Leeds University and distributed by Atkins: www.saturnsoftware.co.uk
SD	Standard deviation
SE	Standard error
SH1 and SH2	State Highways 1 and 2 – main roads in and out of the study area
Slow mode	Walk and cycle
TAM	*Traffic appraisal manual* (UK). Now DMRB, vol 12, section 1, part 1, less some good 1981 bits.
TEC	*Journal of Traffic Engineering & Control*
TLA	Territorial local authority; administrative area smaller than a region
TMfS	Transport Model for Scotland
TN	Technical note: a series documenting the Wellington Transport Strategy Model. Published by Sinclair Knight Merz and Beca Carter Hollings & Ferner (2003)
TRACKS	Transport modelling software: gabites.co.nz/tracks.asp
TransCAD	Transport modelling software, built around a GIS: calliper.com/tcovu.htm
Trip end	Types of trip end: origin – start of trip destination – finish of trip production – home end of trip attraction – non-home end of trip generation – production or attraction and its relation to land use
Trips	Transport modelling software, developed by MVA, now incorporated in Cube
TripsWin	Graphical interface for linking programmes in Trips; now Application Manager in Cube
UTPS	Urban Transportation Planning System
vdf	Volume-delay function, EMME/2 nomenclature
VISUM	Transport modelling software: www.ptvag.com/software/transportation-planning-traffic-engineering/software-system-solutions/visum/
WebTAG	Traffic Analaysis Guidance website of the UK Department for Transport: www.dft.gov.uk/webtag/
Voyager	Transport modelling software – part of the Cube suite
VSN	Suppliers of Genstat
WTSM	Wellington Transport Strategy Model